



UNIVERSIDADE ESTADUAL PAULISTA “JÚLIO DE MESQUITA FILHO”

Instituto de Geociências e Ciências Exatas

Câmpus de Rio Claro

Programa de Pós-Graduação em Ciência da Computação

**Gustavo Furlanetto**

# **X-Net: Uma Arquitetura Bioinspirada para Modulação de Representações Multimodais Através de Sincronização e Reconstrução Cruzada**

**Rio Claro - SP**

**2025**

UNIVERSIDADE ESTADUAL PAULISTA  
“Júlio de Mesquita Filho”  
Instituto de Geociências e Ciências Exatas  
Câmpus de Rio Claro

Gustavo Furlanetto

**X-Net: Uma Arquitetura Bioinspirada  
para Modulação de Representações  
Multimodais Através de Sincronização e  
Reconstrução Cruzada**

Dissertação de Mestrado apresentada ao  
Instituto de Geociências e Ciências Exatas  
do Câmpus de Rio Claro, da Universidade  
Estadual Paulista “Júlio de Mesquita Filho”,  
como parte dos requisitos para obtenção do  
título de Mestre em Ciência da Computação.

Orientador: Prof. Dr. Denis Henrique Pinheiro  
Salvadeo

Rio Claro - SP  
2025

F985x Furlanetto, Gustavo

X-net: uma arquitetura bioinspirada para modulação de representações multimodais através de sincronização e reconstrução cruzada / Gustavo Furlanetto. -- Rio Claro, 2025  
99 p. : il., tabs.

Dissertação (mestrado) - Universidade Estadual Paulista (UNESP), Instituto de Geociências e Ciências Exatas, Rio Claro  
Orientador: Denis Henrique Pinheiro Salvadeo

1. Computer vision. 2. Deep learning (Machine learning). 3. Classificação Multimeios. I. Título.

## **Impacto potencial desta pesquisa**

Esta pesquisa aproxima a inteligência artificial da forma como o ser humano integra visão e audição. Inspirando-se nos processos cerebrais de integração sensorial, propõe estratégias voltadas ao aprimoramento da fusão multimodal entre áudio e vídeo, contribuindo para o desenvolvimento de bases conceituais e tecnológicas que possam tornar a interação homem-máquina mais natural, inclusiva e presente no cotidiano.

## **Potential impact of this research**

This research brings artificial intelligence closer to the way humans integrate vision and hearing. Inspired by the brain's sensory integration processes, it proposes strategies aimed at enhancing multimodal fusion between audio and video, contributing to the development of conceptual and technological foundations that can make human-machine interaction more natural, inclusive, and integrated into everyday life.



UNIVERSIDADE ESTADUAL PAULISTA  
“Júlio de Mesquita Filho”  
Instituto de Geociências e Ciências Exatas  
Câmpus de Rio Claro

Gustavo Furlanetto

**X-Net: Uma Arquitetura Bioinspirada  
para Modulação de Representações  
Multimodais Através de Sincronização e  
Reconstrução Cruzada**

Dissertação de Mestrado apresentada  
ao Instituto de Geociências e Ciências  
Exatas do Câmpus de Rio Claro, da  
Universidade Estadual Paulista “Júlio de  
Mesquita Filho”, como parte dos  
requisitos para obtenção do título de  
Mestre em Ciência da Computação.

Comissão Examinadora

Prof. Dr. Denis Henrique Pinheiro Salvadeo  
IGCE / UNESP/Rio Claro (SP)

Prof. Dr. Hélio Pedrini  
IC/ UNICAMP/Campinas (SP)

Prof. Dr. Rodrigo Capobianco Guido  
IBILCE/ UNESP/Rio Claro (SP)

Conceito: Aprovado.

Rio Claro (SP), 27 de Agosto de 2025.

# Resumo

Os recentes avanços nas pesquisas com modelos multimodais reacenderam o interesse na exploração de novas aplicações na área. Grandes modelos generativos multimodais obtiveram sucesso ao atrair a atenção não apenas de pesquisadores, mas também do público em geral. Inspirados na capacidade humana de lidar com múltiplas modalidades sensoriais, este projeto explorou a convergência entre representações sonoras e visuais, examinando o espaço latente que existe entre elas. Iniciamos nossa investigação com a criação e o treinamento de modelos dedicados à conversão de dados de áudio em imagens e vice-versa. A partir da análise estrutural e representacional desses modelos, aliada ao conhecimento biológico sobre o processamento multimodal no cérebro humano, propusemos uma arquitetura inovadora para o processamento audiovisual: a X-Net. Essa arquitetura é composta por módulos especializados em sincronização temporal e reconstrução cruzada, cuja saída modula dinamicamente os caminhos de processamento unimodal, enriquecendo a extração de características em ambas as modalidades. A X-Net foi avaliada em tarefas de classificação nos conjuntos de dados COIN e UCF101, obtendo melhorias de desempenho consistentes com o uso do módulo de controle multimodal: no COIN, a acurácia Top-1 subiu de 52,3% para 59,1%, e no UCF101, de 75,5% para 81,7%. Esses resultados validam a eficácia da abordagem proposta e reforçam o potencial de mecanismos inspirados biologicamente na construção de modelos multimodais mais interpretáveis e eficientes.

**Palavras-chave:** Processamento Audiovisual. Modelos Multimodais. Visão Computacional. Aprendizado Profundo.

# Abstract

Recent advances in multimodal model research have reignited interest in exploring new applications in the field. Large generative multimodal models have successfully drawn the attention of both researchers and the general public. Inspired by the human ability to integrate multiple sensory modalities, this project explores the convergence between auditory and visual representations by examining the latent space that connects them. The investigation began with the development and training of models dedicated to converting audio into images and vice versa. Based on the structural and representational analysis of these models—and supported by biological knowledge of multimodal processing in the human brain—we proposed X-Net, a bio-inspired architecture for audiovisual multimodal processing. X-Net combines temporal synchronization and cross-modal reconstruction modules, whose outputs are used to modulate unimodal backbones dynamically, enriching the extraction of features in both modalities. The architecture was evaluated on classification tasks using the COIN and UCF101 datasets, achieving consistent performance improvements with the inclusion of the multimodal controller: Top-1 accuracy increased from 52.3% to 59.1% on COIN and from 75.5% to 81.7% on UCF101. These results validate the effectiveness of the proposed approach and highlight the potential of biologically inspired mechanisms in building more interpretable and efficient multimodal models.

**Keywords:** Audiovisual processing. Multimodal models. Computer vision. Deep Learning.

# Lista de Figuras

|     |                                                                         |    |
|-----|-------------------------------------------------------------------------|----|
| 2.1 | Neurônio computacional . . . . .                                        | 19 |
| 2.2 | Rede convolucional . . . . .                                            | 21 |
| 2.3 | Autoencoder . . . . .                                                   | 23 |
| 2.4 | Autoencoder Variacional . . . . .                                       | 25 |
| 2.5 | Redes Generativas Adversariais . . . . .                                | 26 |
| 2.6 | Arquitetura Transformers . . . . .                                      | 28 |
| 2.7 | Geração de um espectrograma . . . . .                                   | 31 |
| 2.8 | Sistema auditivo receptor . . . . .                                     | 32 |
| 2.9 | Sistema visual . . . . .                                                | 34 |
| 3.1 | Processo de treinamento proposto para treinamento da Soundnet . . . . . | 39 |
| 3.2 | Pipeline de geração de áudio . . . . .                                  | 41 |
| 3.3 | Impact Sound Neural Network - AV . . . . .                              | 42 |
| 3.4 | Sonificação da reação do solo . . . . .                                 | 43 |
| 3.5 | Modelo AvSlowFast . . . . .                                             | 46 |
| 3.6 | Modelo AVTS . . . . .                                                   | 47 |
| 3.7 | Estrutura Multisensory. . . . .                                         | 49 |
| 4.1 | Modelos de conversão multimodal . . . . .                               | 52 |
| 4.2 | Imagens dos conjuntos de dados . . . . .                                | 53 |
| 4.3 | Arquitetura da rede usada no treinamento . . . . .                      | 57 |
| 4.4 | Imagens geradas a partir de um conjunto de áudios . . . . .             | 60 |
| 4.5 | Troca de encoders para inferência . . . . .                             | 62 |

---

|      |                                                                                              |    |
|------|----------------------------------------------------------------------------------------------|----|
| 4.6  | Imagens geradas com um VAE misto . . . . .                                                   | 62 |
| 5.1  | Arquitetura da X-Net . . . . .                                                               | 64 |
| 5.2  | YouTube - 8M . . . . .                                                                       | 65 |
| 5.3  | Conjunto UCF101 . . . . .                                                                    | 65 |
| 5.4  | Conjunto COIN . . . . .                                                                      | 66 |
| 5.5  | Arquitetura do ViT aplicada à classificação de imagens . . . . .                             | 67 |
| 5.6  | Processo de pré-treinamento do modelo VideoMAE com mascaramento<br>espaço-temporal . . . . . | 68 |
| 5.7  | Controlador AV . . . . .                                                                     | 69 |
| 5.8  | Arquitetura SincAV . . . . .                                                                 | 71 |
| 5.9  | Arquitetura Processamento AV (Treinamento) . . . . .                                         | 72 |
| 5.10 | Arquitetura Processamento AV (Inferência) . . . . .                                          | 73 |
| 5.11 | Fluxo de informações Controlador AV . . . . .                                                | 76 |
| 5.12 | Treinamento SincAV . . . . .                                                                 | 77 |
| 5.13 | Treinamento Processador AV . . . . .                                                         | 78 |
| 6.1  | Exemplos de reconstrução visual obtidos pelo Processador AV . . . . .                        | 83 |
| 6.2  | Exemplos de reconstrução de espectrogramas obtidos pelo Processador AV                       | 84 |
| 6.3  | Interpolação no espaço latente visual aprendida pelo Processador AV . . .                    | 84 |

# Lista de Tabelas

|     |                                                                         |    |
|-----|-------------------------------------------------------------------------|----|
| 4.1 | Resumo das Configurações Experimentais Avaliadas . . . . .              | 59 |
| 4.2 | Melhor Configuração para os Modelos de Reconstrução . . . . .           | 59 |
| 4.3 | Métricas para conversão de áudio para imagem . . . . .                  | 60 |
| 4.4 | Métricas para conversão de imagem para áudio . . . . .                  | 61 |
| 6.1 | Hiperparâmetros utilizados no treinamento do SincAV . . . . .           | 81 |
| 6.2 | Resultados do Processador AV nas reconstruções unimodais . . . . .      | 82 |
| 6.3 | Hiperparâmetros utilizados no treinamento do Processador AV . . . . .   | 82 |
| 6.4 | Comparação de modelos para a tarefa de classificação de vídeo . . . . . | 85 |
| 6.5 | Hiperparâmetros utilizados no treinamento da X-Net . . . . .            | 86 |

# Lista de Abreviaturas e Siglas

|             |                                     |
|-------------|-------------------------------------|
| <b>BCE</b>  | Binary Cross Entropy                |
| <b>BN</b>   | Batch Normalization                 |
| <b>CNN</b>  | Convolutional Neural Network        |
| <b>FiLM</b> | Feature-wise Linear Modulation      |
| <b>GPU</b>  | Graphics Processing Unit            |
| <b>I/O</b>  | Input/Output                        |
| <b>ML</b>   | Machine Learning                    |
| <b>MMD</b>  | Maximum Mean Discrepancy            |
| <b>MSE</b>  | Mean Squared Error                  |
| <b>ReLU</b> | Rectified Linear Unit               |
| <b>RNN</b>  | Recurrent Neural Network            |
| <b>SSIM</b> | Structural Similarity Index Measure |
| <b>STG</b>  | Giro Temporal Superior              |
| <b>STS</b>  | Sulco Temporal Superior             |
| <b>VAE</b>  | Variational Autoencoder             |

# Sumário

|          |                                                               |           |
|----------|---------------------------------------------------------------|-----------|
| <b>1</b> | <b>Introdução</b>                                             | <b>14</b> |
| 1.1      | Definição do problema . . . . .                               | 14        |
| 1.2      | Motivações e Desafios . . . . .                               | 15        |
| 1.3      | Questão de Pesquisa . . . . .                                 | 15        |
| 1.4      | Objetivos . . . . .                                           | 16        |
| 1.5      | Estrutura do Texto . . . . .                                  | 17        |
| <br>     |                                                               |           |
| <b>2</b> | <b>Fundamentação Teórica</b>                                  | <b>18</b> |
| 2.1      | Aprendizado Profundo . . . . .                                | 18        |
| 2.2      | Redes Neurais Convolucionais (CNN) . . . . .                  | 20        |
| 2.3      | Convoluções 3D . . . . .                                      | 21        |
| 2.4      | Autoencoders . . . . .                                        | 22        |
| 2.5      | Autoencoders Variacionais . . . . .                           | 24        |
| 2.6      | Redes Generativas Adversariais . . . . .                      | 25        |
| 2.7      | Transformers . . . . .                                        | 26        |
| 2.8      | Aprendizado Autossupervisionado . . . . .                     | 28        |
| 2.9      | Representação e Processamento de Dados Audiovisuais . . . . . | 29        |
| 2.10     | Sinais Auditivos . . . . .                                    | 30        |
| 2.10.1   | Estrutura do Sistema Auditivo . . . . .                       | 31        |
| 2.10.2   | Processamento dos sinais elétricos . . . . .                  | 32        |
| 2.11     | Sinais Visuais . . . . .                                      | 33        |
| 2.11.1   | Estrutura do Sistema Visual . . . . .                         | 33        |



---

|          |                                                                    |           |
|----------|--------------------------------------------------------------------|-----------|
| 2.11.2   | Processamento dos sinais elétricos . . . . .                       | 34        |
| 2.12     | Processamento Multimodal . . . . .                                 | 35        |
| 2.12.1   | Regiões Envolvidas no Processamento Multimodal . . . . .           | 35        |
| 2.12.2   | Fluxo de Informações Multimodal . . . . .                          | 36        |
| 2.12.3   | Integração e Percepção . . . . .                                   | 37        |
| <b>3</b> | <b>Trabalhos Relacionados</b>                                      | <b>38</b> |
| 3.1      | Trabalhos do Meio da Computação . . . . .                          | 38        |
| 3.2      | Trabalhos do Meio Biológico . . . . .                              | 41        |
| 3.3      | Classificação de Vídeos . . . . .                                  | 44        |
| <b>4</b> | <b>Um Estudo Exploratório Inicial</b>                              | <b>51</b> |
| 4.1      | Visão Geral . . . . .                                              | 51        |
| 4.2      | Conjuntos de Dados . . . . .                                       | 51        |
| 4.3      | Pré-processamento das Imagens . . . . .                            | 53        |
| 4.4      | Processamento de Áudio e Espectrogramas . . . . .                  | 54        |
| 4.5      | Arquitetura do Modelo . . . . .                                    | 55        |
| 4.6      | Treinamento e Avaliação . . . . .                                  | 57        |
| 4.7      | Metodologia dos Experimentos . . . . .                             | 58        |
| 4.8      | Resultado e Análise do Treinamento . . . . .                       | 60        |
| <b>5</b> | <b>X-Net: Arquitetura de Convergência Audiovisual Bioinspirada</b> | <b>63</b> |
| 5.1      | Visão Geral . . . . .                                              | 63        |
| 5.2      | Conjunto de Dados . . . . .                                        | 64        |
| 5.3      | Arquitetura . . . . .                                              | 66        |
| 5.4      | Controlador AV . . . . .                                           | 68        |
| 5.4.1    | Modulação das Backbones por FiLM . . . . .                         | 74        |
| 5.4.2    | Treinamento da arquitetura . . . . .                               | 76        |
| <b>6</b> | <b>Análise dos Resultados</b>                                      | <b>80</b> |
| 6.1      | Ambiente de Execução . . . . .                                     | 80        |
| 6.2      | Resultados do Módulo SincAV . . . . .                              | 80        |

---

|          |                                        |           |
|----------|----------------------------------------|-----------|
| 6.3      | Resultados do Processador AV . . . . . | 82        |
| 6.4      | Resultados da X-Net . . . . .          | 83        |
| <b>7</b> | <b>Considerações Finais</b>            | <b>87</b> |
|          | <b>Referências</b>                     | <b>90</b> |
|          | <b>DADOS CURRICULARES</b>              | <b>98</b> |

# 1 Introdução

Este capítulo tem como objetivo contextualizar a pesquisa desenvolvida, apresentando os fundamentos, motivações e contribuições que nortearam o trabalho. Inicialmente, discutimos o cenário atual do aprendizado profundo multimodal e sua relevância frente à crescente necessidade de integração entre modalidades sensoriais. Em seguida, destacamos trabalhos correlatos que exploraram a convergência de diferentes fluxos de dados, ressaltando os avanços obtidos e suas limitações. Por fim, são descritos os objetivos centrais e as principais contribuições desta dissertação, situando o leitor quanto ao percurso investigativo que será detalhado nos capítulos seguintes.

## 1.1 Definição do problema

A crescente demanda por sistemas capazes de compreender e processar informações multimodais tem gerado um aumento significativo na pesquisa no aprendizado profundo multimodal (HUANG et al., 2022; GIRDHAR et al., 2023a; SUMMAIRA et al., 2021). À medida que as modalidades de áudio e elementos visuais se tornam cada vez mais interligadas em nosso ambiente digital, surge a necessidade de desenvolver soluções que possam unificar o processamento dessas modalidades de forma eficaz. Nosso projeto busca explorar a fusão das modalidades visual e auditiva para propor uma abordagem que unifique o processamento multimodal.

Em trabalhos anteriores que exploraram a convergência de modalidades de dados, resultados promissores têm sido observados em diversas aplicações. Por exemplo, em pesquisas que unificaram áudio e imagem para o reconhecimento de objetos (AFOURAS et al., 2022), os modelos multimodais demonstraram um aumento significativo no desempenho em relação a abordagens unimodais. Essas abordagens ampliaram as capacidades dos sistemas de visão computacional com informações de áudio (MORRISON et al., 2014; YANG et al., 2022), fornecendo uma compreensão mais profunda de objetos e cenas. Além disso, em domínios de acessibilidade (TAYLOR, 2016; KASHEVNIK et al., 2021), a convergência de áudio e imagem provou ser vital para melhorar a experiência de pessoas com deficiências visuais, possibilitando a descrição de conteúdo visual por meio de informações sonoras.

Outros trabalhos se concentraram na integração de texto e imagem, destacando aplicações notáveis, como a recuperação multimodal de informações (HUANG et al., 2022; APPALARAJU et al., 2023; WANG; JIN; DING, 2022) e a geração de descrições textuais para vídeos (RAMANISHKA et al., 2016; HORI et al., 2017). Modelos multimodais desenvolvidos nesse contexto demonstraram a capacidade de entender e extrair relações semânticas entre texto e imagem, resultando em sistemas de busca mais eficazes e geradores de texto mais precisos. Esses resultados promissores indicam o potencial da convergência de modalidades de dados em melhorar a compreensão de informações e enriquecer as interações homem-máquina em uma variedade de cenários.

## 1.2 Motivações e Desafios

Nossa motivação se baseia na compreensão de que a interação sinérgica entre áudio e elementos visuais é fundamental para uma ampla gama de aplicações. Isso inclui desde a tradução automática de conteúdo de vídeo para pessoas com deficiência auditiva (MOCANU; TAPU, 2021), até a criação de sistemas avançados de legendagem automática para vídeos online (CHE et al., 2017). No entanto, para alcançar esses objetivos, é essencial que os modelos de processamento multimodal compreendam não apenas as informações de cada modalidade, mas também a forma como essas modalidades interagem e se complementam.

A proposta desta pesquisa é intrinsecamente ligada à capacidade inata do ser humano de correlacionar e integrar informações provenientes de diferentes modalidades sensoriais, como a capacidade de imaginar imagens a partir de sons e vice-versa. O embasamento biológico dessa abordagem reside na compreensão de que os sentidos humanos não operam de forma isolada, mas sim em constante interação, resultando em uma experiência sensorial rica e integrada tanto no córtex auditivo quanto no córtex visual (KNÖPFEL et al., 2019; BESLE et al., 2008). Intrinsecamente, os seres humanos possuem a capacidade inata de sintetizar informações auditivas associadas a uma cena que estão observando, fazendo uso de suas experiências passadas para tal processamento. De maneira análoga, somos igualmente proficientes em sintetizar uma representação visual ou cena a partir de estímulos sonoros que estamos ouvindo.

## 1.3 Questão de Pesquisa

Seria viável reproduzir a habilidade biológica de representação multimodal utilizando como base nossos modelos e técnicas atuais? A busca por essa resposta direciona nossa exploração na extração e representação do conhecimento em modelos multimodais para

ambas as modalidades auditivas e visuais. Em particular, os Autoencoders emergem como modelos capazes de extrair características essenciais para o processamento multimodal. Nesse contexto, realizamos uma análise do comportamento desses modelos quando empregados no processamento multimodal de dados, com foco na geração eficaz de informações audiovisuais.

Para conseguirmos elaborar a nova abordagem, primeiramente criamos e treinamos dois Autoencoders Variacionais (KINGMA; WELLING, 2013; TOLSTIKHIN et al., 2017) para geração de imagens a partir de uma faixa de áudio e vice-versa. A análise dos modelos treinados evidenciou uma forte correlação intrínseca entre as modalidades, permitindo a geração de dados consistentes mesmo quando os *encoders* dos modelos foram invertidos. Essa análise proporcionou um entendimento mais aprofundado das inter-relações entre as modalidades, o que possibilitou a proposição de uma abordagem unificada, denominada X-Net.

A nova arquitetura propõe o *Controlador AV*, um bloco central de processamento audiovisual inspirado em processos biológicos humanos de integração multimodal (BEAUCHAMP et al., 2004; SCHEEF et al., 2009). De maneira análoga ao processamento realizado pelo cérebro humano, esse bloco, treinado integralmente com técnicas de aprendizado autossupervisionado, realiza o processamento conjunto das modalidades, avaliando a sincronicidade dos dados e explorando a capacidade de conversão entre modalidades distintas, conforme identificado em nossa análise dos modelos de conversão.

O *Controlador AV* é integrado a duas *backbones* unimodais tradicionais, com o objetivo de complementar as informações extraídas por essas estruturas. Essa integração estabelece um paralelo direto com a organização biológica do cérebro humano, reforçando a inspiração biológica da proposta. Esse método se destaca em relação a outras abordagens multimodais por propor o cruzamento entre modalidades no *Controlador AV*, além de realizar a verificação da sincronicidade entre áudio e vídeo. Por fim, conduzimos uma análise comparativa do desempenho da nova arquitetura em relação a trabalhos correlatos que utilizam estímulos visuais ou auditivos no contexto do processamento multimodal de vídeos.

## 1.4 Objetivos

O principal objetivo deste trabalho é investigar como mecanismos inspirados na cognição humana podem ser aplicados na construção de arquiteturas neurais capazes de integrar diferentes modalidades sensoriais de forma eficiente. Especificamente, buscamos projetar uma solução multimodal para tarefas de classificação de vídeos que vá além da simples concatenação de *features*, incorporando elementos de sincronização temporal, reconstrução

cruzada e modulação condicional entre os fluxos auditivo e visual. A proposta visa não apenas melhorar o desempenho em tarefas supervisionadas, mas também explorar técnicas de aprendizado autosupervisionado, reduzindo a dependência de grandes volumes de dados anotados.

Como principais contribuições, destacamos: (i) a criação de uma arquitetura multi-modal bioinspirada, composta por módulos especializados que emulam processos neurais de sincronização e integração sensorial; (ii) o desenvolvimento do módulo *SincAV*, capaz de identificar alinhamento temporal entre áudio e vídeo em regime autosupervisionado; (iii) a proposta de um módulo de reconstrução cruzada entre modalidades, com base em *autoencoders variacionais*, que permite o aprendizado de representações latentes compartilhadas; e (iv) a incorporação de um mecanismo de modulação via *Feature-wise Linear Modulation* (Feature-wise Linear Modulation (FiLM)), que condiciona dinamicamente as *backbones* unimodais com informações extraídas do fluxo complementar. Essa combinação de estratégias resulta em uma arquitetura modular, interpretável e eficiente, avaliada de forma abrangente em múltiplos conjuntos de dados audiovisuais.

## 1.5 Estrutura do Texto

Nos próximos capítulos, iremos explorar a nossa proposta em diferentes níveis. O Capítulo 2 deste trabalho concentra-se na construção de uma base robusta para nossa proposta, oferecendo uma explanação detalhada dos fundamentos teóricos que permeiam os conceitos centrais envolvidos. Já no Capítulo 3, realizamos uma análise aprofundada de estudos relacionados. Essa revisão não só fornece uma fonte valiosa de motivação para nossa pesquisa, mas também estabelece um conjunto de referências que servirão como parâmetros para comparações com a abordagem que apresentamos. Em seguida, no Capítulo 4, exploramos de maneira aprofundada todos os elementos da nossa pesquisa sobre conversão audiovisual, abrangendo aspectos como a arquitetura dos modelos, o processamento dos dados, experimentos conduzidos e análise dos resultados. No Capítulo 5, utilizamos os aprendizados obtidos com os modelos de conversão para propor a X-Net. Expomos aqui todos os aspectos envolvidos para a criação dessa nova arquitetura, incluindo conjunto de dados, análise das estruturas, formas de treinamento e experimentos pretendidos. No Capítulo 6, realizamos uma análise aprofundada dos resultados obtidos. Por fim, no Capítulo 7, são apresentadas a conclusão geral e uma análise crítica dos principais resultados obtidos, além de uma reflexão sobre possíveis direções para trabalhos futuros, seja pela exploração de novas abordagens ou pela ampliação do presente estudo.

## 2 Fundamentação Teórica

Neste capítulo, iremos elucidar a fundamentação teórica da nossa proposta. Aprofundaremos nossa análise nos conceitos fundamentais do aprendizado profundo, expondo as técnicas e arquiteturas que permeiam esse campo dinâmico. Além disso, exploraremos as representações e métodos essenciais no âmbito do processamento de áudio e vídeo digitais.

### 2.1 Aprendizado Profundo

O Aprendizado Profundo é uma subárea do campo de Aprendizado de Máquina (em inglês, *Machine Learning* - Machine Learning (ML)) que busca emular o processo de aprendizado humano através do treinamento de redes neurais artificiais profundas. A base do aprendizado profundo está na capacidade dessas redes em aprender representações cada vez mais abstratas a medida que o processamento avança nas camadas (LI et al., 2013).

Apesar do ponto inicial do campo estar entre 1960 e 1970, com o desenvolvimento das primeiras redes neurais, a eficácia dessas redes só foi plenamente realizada mais tarde, com o aumento significativo na quantidade de dados disponíveis e o avanço das Unidade de Processamento Gráfico (em inglês, *Graphics Processing Unit* – Graphics Processing Unit (GPU)). A popularização de arquiteturas como as Redes Neurais Convolucionais (em inglês, *Convolutional Neural Networks* – Convolutional Neural Network (CNN)), as Redes Neurais Recorrentes (em inglês, *Recurrent Neural Networks* – Recurrent Neural Network (RNN)), e mais recentemente os *transformers*, impulsionou o crescimento e sucesso do aprendizado profundo em diversas aplicações (GELLY; GAUVAIN, 2017).

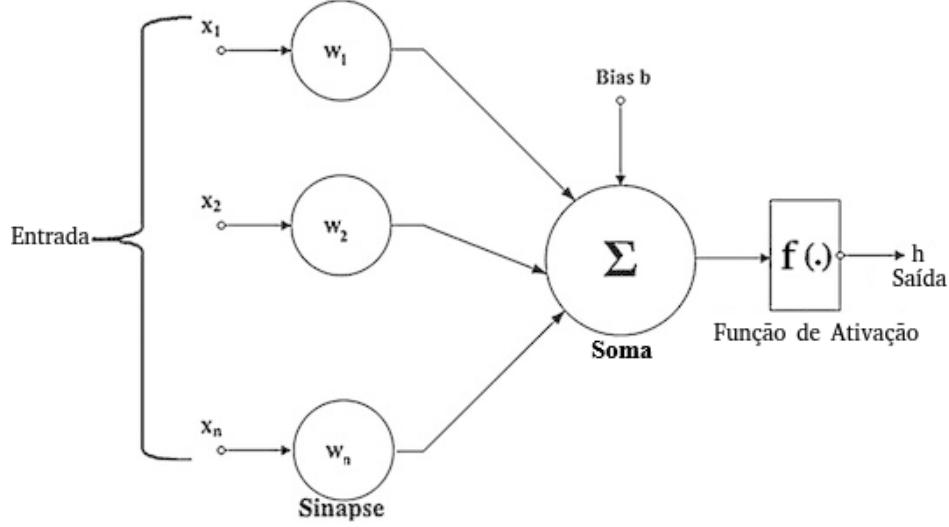
A representação do neurônio fundamental muito utilizado no campo de aprendizado profundo, observado na Figura 2.1, é a unidade básica de processamento em uma rede neural. Inspirado no funcionamento dos neurônios biológicos, um neurônio artificial recebe entradas  $x_i$ , aplica uma função de ativação  $f$  e produz uma saída  $h$ :

$$h = f\left(\sum_i w_i x_i + b\right). \quad (2.1)$$

Aqui,  $w_i$  são os pesos associados às entradas  $x_i$ , e  $b$  é o viés (*bias*). Os pesos controlam a contribuição de cada entrada na saída, e o *bias* permite ao neurônio ajustar seu

comportamento para algum viés detectado.

Figura 2.1: Neurônio computacional



Fonte: Adaptado de Academy (2022).

Dada a sua versatilidade, modelos de aprendizado profundo são amplamente utilizados para uma gama de aplicações. A capacidade de aprendizado hierárquico permite que esses modelos extraiam de forma automática características relevantes, eliminando a necessidade de pré-processamento intensivo. Além disso, o aprendizado profundo destaca-se em tarefas como reconhecimento de padrões em imagens (FU, 1976), tradução automática (BÉRARD et al., 2018) e processamento de linguagem natural (NADKARNI; OHNO-MACHADO; CHAPMAN, 2011).

O aprendizado em redes profundas é regido pelo processo de otimização de parâmetros com base em um conjunto de dados de treinamento. Seja um modelo representado por uma função  $f(\theta)$ , em que  $\theta$  denota o vetor de parâmetros do modelo. O objetivo é minimizar uma função de perda  $L(\theta)$ , que mede a diferença entre os resultados obtidos pelo modelo e os rótulos pré-determinados do conjunto de treinamento. Durante o processo iterativo de treinamento, indicamos por  $\theta^{(t)}$  o conjunto de parâmetros na iteração  $t$ , e por  $\theta^{(t+1)}$  os parâmetros atualizados após a aplicação de uma etapa de descida do gradiente:

$$\theta^{(t+1)} = \theta^{(t)} - \alpha \nabla L(\theta^{(t)}), \quad (2.2)$$

em que  $\alpha$  representa a taxa de aprendizado e  $\nabla L(\theta^{(t)})$  é o gradiente da função de perda em relação aos parâmetros na iteração  $t$ . Esse gradiente indica a direção de maior crescimento da função de perda, e a taxa de aprendizado controla o tamanho dos passos de atualização.

O aprendizado em redes profundas é composto pela iteração contínua sobre esse processo, permitindo que o modelo ajuste gradualmente seus pesos para otimizar o resultado em



relação ao conjunto de treinamento. Esse ajuste contínuo permite a rede a aprender representações cada vez mais complexas e hierárquicas, tornando-se capaz de realizar tarefas específicas com maior precisão.

## 2.2 Redes Neurais Convolucionais (CNN)

As Redes Neurais Convolucionais (em inglês, Convolutional Neural Network - CNNs) são redes de aprendizado profundo, desenvolvidas especificamente para lidar com dados matriciais, como imagens (LI et al., 2014; KRIZHEVSKY; SUTSKEVER; HINTON, 2012). Sua estrutura é inspirada na organização visual do córtex cerebral, onde neurônios individuais respondem a regiões específicas do campo visual.

Uma CNN é composta por camadas convolucionais, de *pooling* e completamente conectadas. A camada convolucional é responsável pela detecção de padrões nas imagens, utilizando filtros que percorrem a entrada e aplicam a função de convolução (KRIZHEVSKY; SUTSKEVER; HINTON, 2012). A operação de convolução aplica um filtro (também chamado de *kernel*) sobre a imagem de entrada para extrair características relevantes. Dada uma imagem  $I$  e um *kernel*  $K$ :

$$S(i, j) = (I * K)(i, j) = \sum_m \sum_n I(m, n) \cdot K(i - m, j - n). \quad (2.3)$$

Aqui,  $S(i, j)$  é o valor resultante após aplicar o filtro na posição  $(i, j)$  da imagem. A operação de convolução é realizada deslizando o *kernel* sobre a imagem e multiplicando e somando os elementos correspondentes (LI et al., 2014).

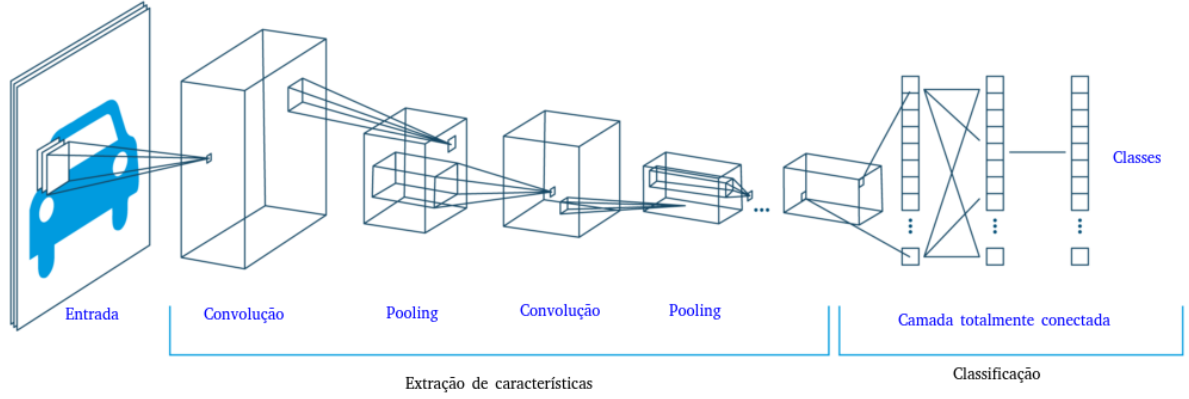
A operação de *pooling* é comumente empregada em redes convolucionais para reduzir a dimensionalidade espacial dos mapas de características, tentando preservar as características mais importantes. Há diversos tipos de *pooling*, o mais comum é o *max pooling*, que seleciona o valor máximo de um conjunto de valores em uma região  $R$  específica:

$$P(i, j) = \max_{(m, n) \in R} I(i + m, j + n), \quad (2.4)$$

onde  $P(i, j)$  é o valor resultante após aplicar a operação de *max pooling*,  $I(i + m, j + n)$  representa os elementos na região  $R$ , e  $(m, n)$  varre essa região com um passo  $s$  (KRIZHEVSKY; SUTSKEVER; HINTON, 2012). Essa técnica é eficaz para reduzir a complexidade computacional e melhorar a invariância a pequenas translações no espaço. Como as rede de aprendizado profundo podem se tornar excessivamente grandes, para segurar o custo computacional e temporal do processamento podemos utilizar a técnica de *pooling* (KRIZHEVSKY; SUTSKEVER; HINTON, 2012). Já as camadas completamente

conectadas finalizam a rede. Nessas camadas, cada neurônio está conectado a todos os neurônios da camada anterior, formando representações mais abstratas e combinando informações de várias regiões da entrada.

Figura 2.2: Rede convolucional



Fonte: Adaptado de Vidhya (2021b)

CNNs, como a observada na Figura 2.2, revolucionaram diversas aplicações, notavelmente no reconhecimento de padrões visuais. Elas são amplamente utilizadas em tarefas como classificação de imagens (LI et al., 2014), detecção de objetos (ŠEVO; AVRAMOVIĆ, 2016), segmentação semântica (MOESKOPS et al., 2016). A versatilidade das CNNs as torna essenciais em projetos que envolvem processamento de imagem.

## 2.3 Convoluções 3D

As convoluções tridimensionais (3D) são uma extensão natural das convoluções bidimensionais (2D), adaptadas para processar dados com estrutura temporal ou volumétrica, como vídeos ou imagens médicas tridimensionais (TRAN et al., 2015). Enquanto as CNNs tradicionais aplicam filtros em duas dimensões espaciais (altura e largura), as convoluções 3D incorporam também a dimensão temporal (ou de profundidade), permitindo a extração de características que variam ao longo do tempo ou no volume (JI et al., 2012; TRAN et al., 2015). Em uma convolução 3D, o filtro (ou *kernel*) é tridimensional e percorre o volume de entrada em três eixos: altura, largura e profundidade/tempo. A operação pode ser definida da seguinte forma. Dado um volume de entrada  $V$  e um filtro tridimensional  $K$  (JI et al., 2012):

$$S(i, j, k) = (V * K)(i, j, k) = \sum_m \sum_n \sum_p V(m, n, p) \cdot K(i - m, j - n, k - p). \quad (2.5)$$

Neste caso,  $S(i, j, k)$  representa o valor resultante da aplicação do filtro na posição  $(i, j, k)$  do volume de entrada (JI et al., 2012). O filtro é deslizado tridimensionalmente, realizando multiplicações e somas que capturam padrões locais não apenas no espaço, mas também no tempo ou profundidade.

Assim como nas CNNs convencionais, as redes com convoluções 3D também empregam camadas de *pooling* tridimensional. O *3D max pooling*, por exemplo, reduz a dimensionalidade espacial e temporal ao selecionar o valor máximo dentro de uma região tridimensional  $R$  (BHATTAD et al., 2023):

$$P(i, j, k) = \max_{(m, n, p) \in R} V(i + m, j + n, k + p). \quad (2.6)$$

Esse processo é fundamental para reduzir a complexidade computacional e evitar sobreajuste, ao mesmo tempo em que mantém as informações mais relevantes da entrada.

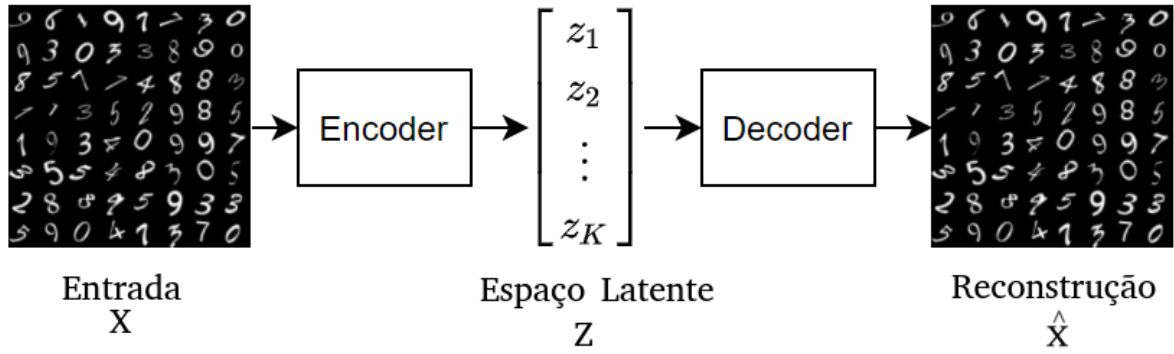
As convoluções 3D são amplamente utilizadas em tarefas que envolvem informação temporal, como classificação de ações em vídeos, reconhecimento de atividades e segmentação de vídeo (JI et al., 2012; TRAN et al., 2015; BHATTAD et al., 2023). Ao processar sequências de quadros como um volume contínuo, essas redes são capazes de capturar padrões dinâmicos, como movimentos, transições e interações espaço-temporais, que não podem ser representados eficientemente com convoluções 2D. Em contextos onde o entendimento da evolução no tempo é essencial, as convoluções 3D tornam-se uma ferramenta indispensável.

## 2.4 Autoencoders

Os Autoencoders são modelos de aprendizado que buscam aprender uma representação compacta e eficiente dos dados de entrada. Como pode ser observado na Figura 2.3, Autoencoders consistem em duas partes principais: o codificador (*encoder*), que converte os dados de entrada em uma representação vetorial (latente), e o decodificador (*decoder*), que reconstrói os dados a partir dessa representação. A arquitetura dos Autoencoders pode variar, mas geralmente é simétrica (KINGMA; WELLING, 2013; MATHWORKS, 2024).

A função do *encoder* é mapear a entrada original para uma representação latente de dimensionalidade inferior (KINGMA; WELLING, 2013). Isso é realizado por meio de camadas de transformação, onde os pesos e os vieses são ajustados durante o treinamento para aprender uma representação informativa da entrada. Matematicamente, o *encoder* pode ser representado por uma função  $f_{\text{encoder}}$  que mapeia a entrada  $x$  para a representação latente  $z$ :

Figura 2.3: Autoencoder



Fonte: Adaptado de MathWorks (2024)

$$z = f_{\text{encoder}}(x). \quad (2.7)$$

Aqui,  $z$  é a representação latente de  $x$ , que captura as características mais essenciais da entrada (KINGMA; WELLING, 2013). Para um *encoder* baseado em camadas convolucionais,  $f_{\text{encoder}}$  assume a equação das convoluções exposta na Equação 2.3.

O *decoder* tem a função de gerar uma reconstrução  $\hat{x}$  a partir da representação latente  $z$ . Seja  $f_{\text{decoder}}$  a função de decodificação, a reconstrução pode ser expressa como:

$$\hat{x} = f_{\text{decoder}}(z), \quad (2.8)$$

onde  $z$  representa a representação latente aprendida pelo *encoder* (KINGMA; WELLING, 2013; MATHWORKS, 2024). Essa operação visa reproduzir a entrada original  $x$  da melhor forma possível. Podemos também usar o *decoder* para gerar um dado novo ou modificado, nesse caso,  $\hat{x}$  assume a forma dessa modificação desejada. Analogamente ao  $f_{\text{encoder}}$ , o  $f_{\text{decoder}}$  também pode assumir a forma das convoluções, mas nesse caso convoluções transpostas, também conhecidas como *deconvolutions* (MATHWORKS, 2024). Esse tipo de camada é uma operação crucial nos *decoders* baseados em CNNs, como nos *Autoencoders* e GANs. Essa operação busca a expansão de características aprendidas no espaço latente para uma representação mais detalhada e dimensionalmente maior.

Enquanto uma convolução padrão reduz a dimensão espacial de uma entrada (*down-sampling*), as convoluções transpostas fazem o oposto, aumentando a dimensão espacial (*upsampling*) (YOUNESI et al., 2024). Elas são utilizadas para criar mapas de características mais densos, restaurando detalhes e estruturas durante o processo de codificação. O processo de *upsampling* também é aprendido durante o treinamento.

Convoluções transpostas podem ser representadas pela operação de convolução conven-

cional, mas com os papéis de entrada e saída trocados. Suponha que  $\mathbf{X}$  seja a entrada,  $\mathbf{W}$  seja o filtro e  $\mathbf{Y}$  seja a saída da convolução transposta (YOUNESI et al., 2024):

$$\mathbf{Y}[i, j] = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} \mathbf{X}[i + m, j + n] \cdot \mathbf{W}[m, n]. \quad (2.9)$$

Existem várias variantes de Autoencoders, cada uma adaptada a diferentes tipos de dados ou objetivos. Os Autoencoders Variacionais (em inglês, Variational Autoencoder - Variational Autoencoder (VAE)s), observado na Figura 2.4, introduzem uma componente estocástica, gerando representações mais suaves e possibilitando a geração de novas amostras. Já os Denoising Autoencoders (DAEs) são treinados para reconstruir dados corrompidos (VINCENT et al., 2010).

Os Autoencoders encontram aplicação em diversas áreas, como geração de imagens (JORDON; YOON; SCHAAR, 2018), remoção de ruídos (TRAN; NGUYEN; ARAI, 2020) e até transferência de estilo (AZADI et al., 2018). Em particular, os VAEs têm se destacado na geração de dados realistas e na exploração do espaço latente. Por esse motivo, escolhemos utilizar essa arquitetura como base da nossa pesquisa

## 2.5 Autoencoders Variacionais

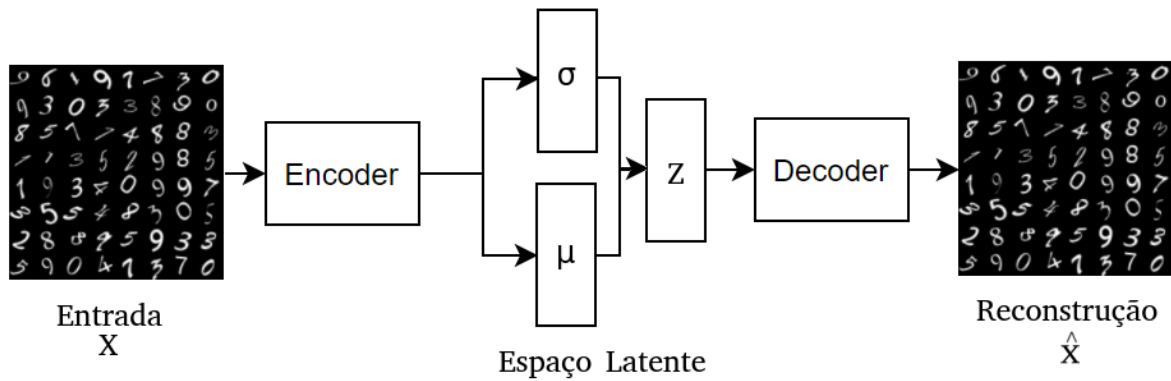
Autoencoders Variacionais introduzem uma abordagem mais probabilística para a representação latente quando comparado ao autoencoder convencional. Para as VAEs, assumimos que a latência  $z$  segue uma distribuição probabilística, geralmente uma distribuição normal multivariada (FENG et al., 2021; MAHMUD et al., 2025). Essa abordagem permite uma maior flexibilidade na geração de novas amostras durante o processo de decodificação. A fórmula geral para uma VAE pode ser expressa da seguinte forma:

$$q_\phi(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2 \mathbf{I}), \quad (2.10)$$

onde  $x$  é a entrada,  $\boldsymbol{\mu}$  e  $\boldsymbol{\sigma}$  são parâmetros aprendidos pelo modelo, e  $I$  é a matriz identidade (FENG et al., 2021). Essa distribuição condicional reflete a variabilidade associada à representação latente dada uma entrada específica, permitindo a amostragem de diferentes instâncias na geração de dados (MAHMUD et al., 2025). Esse aspecto probabilístico é crucial para o treinamento e a geração eficaz de novas amostras em VAEs.

VAEs são especialmente úteis em tarefas de geração de dados, como a reconstrução de imagens ou áudios, devido à natureza probabilística da latência. Isso proporciona uma incerteza inerente nas amostras geradas, permitindo uma maior diversidade e controle sobre os resultados (FENG et al., 2021; MAHMUD et al., 2025).

Figura 2.4: Autoencoder Variacional



Fonte: Composição do Autor

Outra aplicação notável dos VAEs é na realização de tarefas de transferência de domínio e aprendizado semi-supervisionado (FENG et al., 2021). A capacidade de modelar a (KINGMA; WELLING, 2013) distribuição latente dos dados torna os VAEs eficazes em aprender representações compactas e generalizadas, o que é fundamental para melhorar o desempenho em cenários onde os dados são escassos ou as condições de treinamento variam.

Em resumo, as aplicações dos VAEs se destacam em contextos onde é crucial modelar a distribuição probabilística dos dados e obter representações latentes ricas. Essa flexibilidade faz dos VAEs uma escolha poderosa em uma variedade de cenários, desde a geração criativa de dados, como no nosso caso, até a resolução de problemas em que a incerteza e a generalização são fatores-chave.

## 2.6 Redes Generativas Adversariais

Tangenciando o nosso projeto no âmbito de geração de dados, as Redes Generativas Adversariais (GANs) são notadamente reconhecidas por sua capacidade de gerar dados com alta fidelidade (GOODFELLOW et al., 2014; BAU et al., 2018). Essa classe de modelos consistem em duas redes neurais distintas: o gerador ( $G$ ) e o discriminador ( $D$ ). Essas redes são treinadas simultaneamente, envolvendo uma competição entre elas. O objetivo principal é permitir que o gerador aprenda a criar amostras que se assemelhem o máximo possível a dados reais, enquanto o discriminador aprimora sua capacidade de distinguir entre dados genuínos e sintéticos (GOODFELLOW et al., 2014).

A formulação básica de uma GAN é fundamentada em uma função de perda que reflete o jogo entre o gerador e o discriminador (GOODFELLOW et al., 2014). O termo  $L_{GAN}(D, G)$  representa a função de perda total da GAN, que é maximizada pelo discriminador para

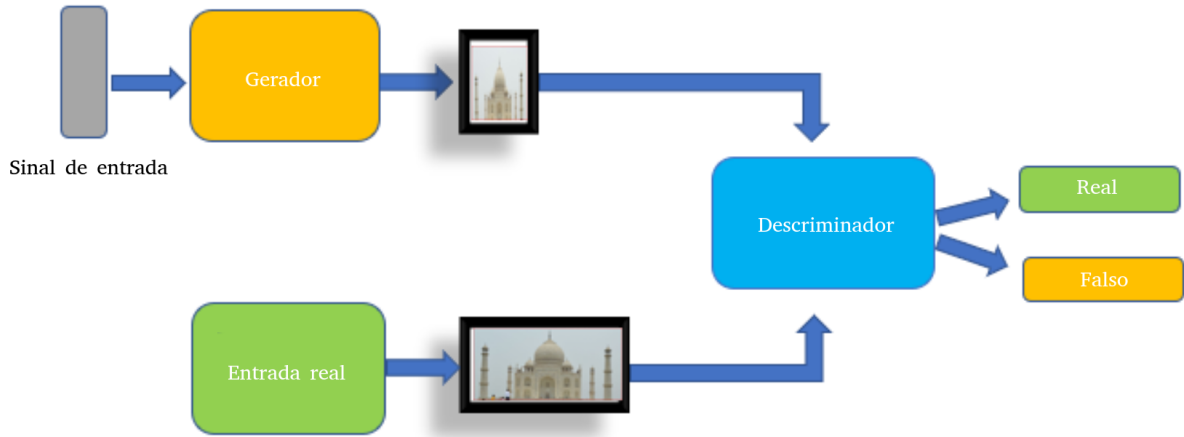
corretamente classificar amostras reais e sintéticas, e ao mesmo tempo, é minimizada pelo gerador para criar amostras que enganem o discriminador:

$$\mathcal{L}_{GAN}(D, G) = \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}(\mathbf{x})}[\log D(\mathbf{x})] + \mathbb{E}_{\mathbf{z} \sim p_{\text{noise}}(\mathbf{z})}[\log(1 - D(G(\mathbf{z})))] \quad (2.11)$$

Aqui,  $x$  representa amostras reais,  $z$  são amostras de um espaço latente  $p_z(z)$ ,  $G(z)$  são as amostras geradas pelo gerador, e  $D(x)$  é a probabilidade atribuída pelo discriminador à amostra  $x$  de ser real (BAU et al., 2018).

Durante o treinamento, os parâmetros de  $G$  são ajustados para minimizar  $L_{GAN}(D, G)$ , enquanto os parâmetros de  $D$  são ajustados para maximizar  $L_{GAN}(D, G)$  (GOODFELLOW et al., 2014). Esse processo iterativo, observado na Figura 2.5, continua até que um ponto de equilíbrio seja alcançado, onde o gerador é capaz de criar amostras indistinguíveis de dados reais, e o discriminador é eficaz em discernir entre amostras autênticas e sintéticas.

Figura 2.5: Redes Generativas Adversariais



Fonte: Adaptado de Vidhya (2021a)

GANs foram inicialmente introduzidas para criar amostras visualmente convincentes, como imagens de rostos humanos, paisagens e até mesmo obras de arte (JIANG; SWEET-SER, 2022). Além disso, as GANs têm sido empregadas com sucesso em tarefas de tradução de estilo (BEJIGA; MELGANI, 2018), onde são capazes de transferir características estilísticas de uma imagem para outra.

## 2.7 Transformers

Os *Transformers* representam uma das mais significativas inovações em redes neurais nos últimos anos, tendo revolucionado o campo de processamento de linguagem natural

(PLN) e, mais recentemente, se estendido para outras áreas como visão computacional e multimodalidade (VASWANI, 2017). Os *Transformers* introduziram uma nova arquitetura, observada na Figura 2.6, que se distancia das tradicionais redes recorrentes e convolucionais, priorizando o uso de mecanismos de atenção para capturar dependências em sequências de dados.

O componente central do *Transformer* é o mecanismo de atenção, mais especificamente a atenção *self-attention* (VASWANI, 2017). Este mecanismo permite que o modelo avalie a importância relativa de cada elemento da sequência em relação a todos os outros elementos, capturando dependências de curto e longo alcance de maneira eficiente. Formalmente, dado um conjunto de vetores de entrada  $X = [x_1, x_2, \dots, x_n]$  (VASWANI, 2017), a atenção é calculada usando as fórmulas:

$$\text{Attention}(Q, K, V) = \text{softmax} \left( \frac{QK^T}{\sqrt{d_k}} \right) V \quad (2.12)$$

onde  $Q$  (*queries*),  $K$  (*keys*) e  $V$  (*values*) são projeções lineares do vetor de entrada  $X$ , e  $d_k$  é a dimensionalidade dos vetores de *keys* (VASWANI, 2017). A aplicação da função *softmax* garante que os pesos atribuídos a cada elemento na sequência somem 1, funcionando como uma medida de importância relativa.

A arquitetura do *Transformer* é composta por dois blocos principais: o *encoder* e o *decoder*. Ambos são formados por múltiplas camadas idênticas empilhadas, onde cada camada consiste de dois sub-blocos: um mecanismo de atenção multi-cabeça (*multi-head self-attention*) e uma rede feed-forward totalmente conectada (VASWANI, 2017).

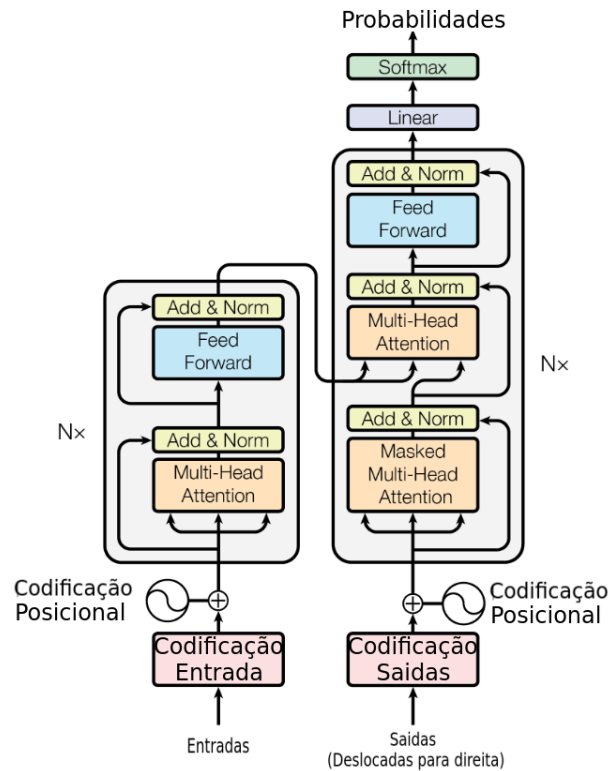
O *encoder* do *Transformer* recebe uma sequência de entrada e aplica sucessivamente os sub-blocos de atenção e feed-forward (VASWANI, 2017). O mecanismo de atenção multi-cabeça é crucial, pois permite ao modelo focar em diferentes partes da sequência simultaneamente, aprendendo diferentes representações contextuais. Este componente é seguido por uma camada de normalização e por uma rede feed-forward que é aplicada de forma independente a cada posição (VASWANI, 2017).

O *decoder* funciona de maneira semelhante ao *encoder*, com a diferença de que incorpora informações da sequência de entrada por meio de uma camada adicional de atenção sobre a saída do *encoder* (VASWANI, 2017). Isso garante que o *decoder* gere saídas que estejam condicionadas não apenas nas posições anteriores da sequência de saída, mas também nas informações processadas pelo *encoder* (VASWANI, 2017). Este mecanismo de atenção cruzada é essencial para tarefas como tradução automática, onde a sequência de entrada (texto na língua de origem) guia a geração da sequência de saída (texto na língua de destino).

Desde sua introdução, os *Transformers* têm sido adotados em uma variedade de



Figura 2.6: Arquitetura Transformers



Fonte: Adaptado de Vaswani (2017)

tarefas além do PLN. Na visão computacional, modelos como o Vision Transformer (ViT) (DOSOVITSKIY et al., 2020) adaptaram a arquitetura dos *Transformers* para lidar com imagens, dividindo-as em *patches* que são tratados como palavras em uma sequência. Em tarefas multimodais, como aquelas envolvendo texto e imagem, os *Transformers* também têm mostrado excelente desempenho, integrando informações de diferentes modalidades em um único modelo (PARTHASARATHY; SUNDARAM, 2021).

Além disso, modelos baseados em *Transformers*, como o BERT (DEVLIN, 2018) e o GPT (BROWN, 2020), têm redefinido o estado da arte em uma série de tarefas de PLN, como classificação de texto, resposta a perguntas, e geração de texto. A flexibilidade dos *Transformers* em lidar com sequências de comprimento variável e a capacidade de capturar dependências complexas entre elementos da sequência têm sido a chave para seu sucesso.

## 2.8 Aprendizado Autossupervisionado

O *Self-Supervised Learning* (SSL), ou Aprendizado autossupervisionado, é uma abordagem emergente no campo de aprendizado de máquina, especialmente em cenários onde há uma escassez de dados rotulados (MISRA; MAATEN, 2020; KRISHNAN; RAJPURKAR;

TOPOL, 2022). Diferentemente dos métodos tradicionais de aprendizado supervisionado, onde a aprendizagem depende de grandes quantidades de dados rotulados, o SSL busca explorar dados não rotulados para criar tarefas preditivas auxiliares, permitindo que o modelo aprenda representações úteis que podem ser posteriormente transferidas para tarefas supervisionadas com dados rotulados.

Por exemplo, no domínio da visão computacional, uma tarefa pretextual comum é a predição de partes ausentes de uma imagem, como preencher um quebra-cabeça, prever a rotação correta de uma imagem, ou até mesmo gerar cores a partir de uma imagem em escala de cinza. Já no domínio de processamento de linguagem natural, uma tarefa pretextual típica pode ser a predição de palavras ocultas em uma sentença, como realizado nos modelos *BERT* (DEVLIN, 2018).

Embora o SSL tenha alcançado resultados promissores, ainda existem desafios significativos associados a essa abordagem. Um dos principais desafios é a criação de tarefas pretextuais eficazes que capturem verdadeiramente as propriedades desejadas dos dados. Além disso, o ajuste fino dos modelos pré-treinados via SSL em tarefas específicas nem sempre resulta em ganhos de desempenho, dependendo da proximidade da tarefa pretextual com a tarefa final.

Outro desafio reside na escalabilidade e eficiência computacional. Modelos SSL frequentemente exigem grande poder de processamento e tempo de treinamento, especialmente quando aplicados a grandes volumes de dados. Técnicas como aprendizado por contraste (*contrastive learning*) (JAISWAL et al., 2020) têm sido exploradas para mitigar esses desafios, mas ainda há espaço para avanços significativos.

## 2.9 Representação e Processamento de Dados Audiovisuais

A estrutura básica de uma imagem é composta por pixels, que são os elementos discretos que armazenam informações sobre cor e intensidade luminosa. Em imagens coloridas, geralmente utiliza-se o modelo RGB (Red, Green, Blue), onde cada pixel é representado por três valores correspondentes às intensidades dessas cores primárias. Essa representação permite a construção de matrizes bidimensionais, onde cada entrada corresponde a um pixel na imagem (TIAN et al., 2019).

A aplicação de técnicas de processamento de imagens, combinada com o poder computacional das CNNs, possibilita a extração de representações hierárquicas complexas, permitindo a detecção de padrões e objetos em diferentes níveis de abstração.

Em relação ao processamento de áudio, aplicamos uma série de transformações sobre o

sinal para que possamos trabalhar com técnicas de tratamento de imagem também, sendo a transformada de Fourier discreta (DFT) uma ferramenta fundamental nesse contexto (TIAN et al., 2019). A DFT permite a decomposição de um sinal de áudio no domínio das frequências, revelando as diferentes componentes espectrais presentes. A fórmula da DFT é dada por:

$$X[k] = \sum_{n=0}^{N-1} x[n] \cdot e^{-j \cdot 2\pi \cdot k \cdot \frac{n}{N}}, \quad (2.13)$$

onde  $X[k]$  representa a amplitude da componente de frequência  $k$ ,  $x[n]$  é a amostra do sinal de áudio no instante  $n$ ,  $N$  é o número total de amostras (TIAN et al., 2019).

Além da DFT, podemos ainda buscar a representação do áudio como um espectrograma, observado na Figura 2.6, que fornece uma visão tanto sobre as frequências quanto informações temporais do conteúdo de um sinal ao longo do tempo. O espectrograma é obtido aplicando repetidamente a DFT em janelas temporais do sinal, denominada Transformada de Fourier de curta duração (Short-Time Fourier Transform - STFT) (STURMEL; DAUDET et al., 2011; TIAN et al., 2019). Denotando o sinal de áudio como  $x[n]$  e escolhendo uma janela de análise  $w[n]$ , o espectrograma  $S[k, m]$  pode ser calculado como:

$$S[k, m] = \sum_{n=0}^{N-1} x[n + mH] \cdot w[n] \cdot e^{-j \cdot 2\pi \cdot k \cdot \frac{n}{N}}, \quad (2.14)$$

onde  $H$  é o passo de avanço (ou salto) entre as janelas sucessivas e  $m$  indica o índice da janela (STURMEL; DAUDET et al., 2011).

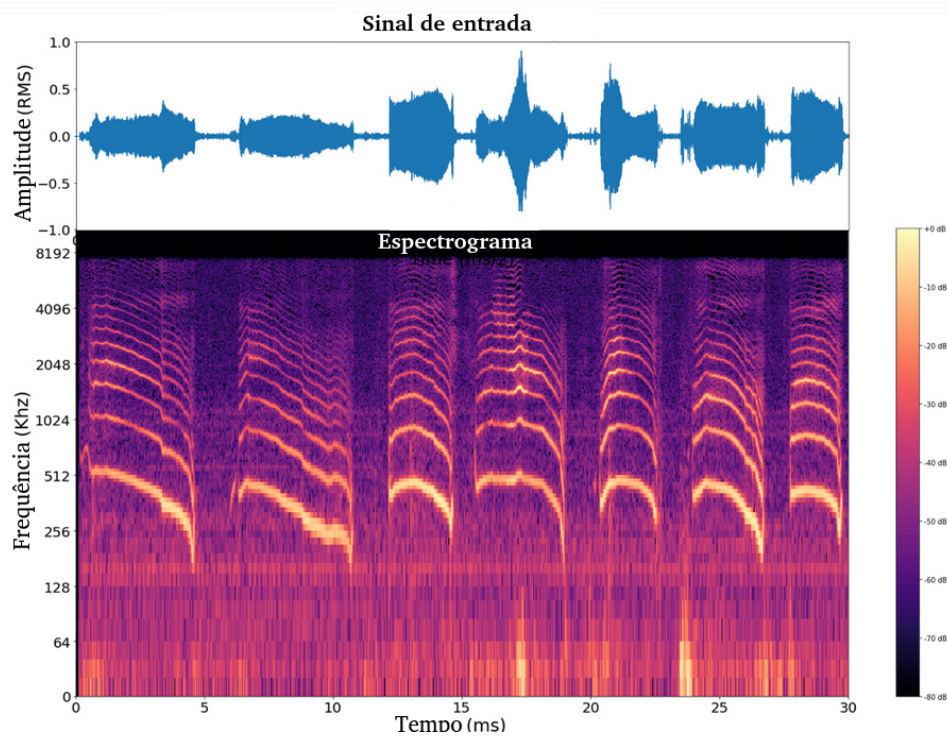
Essa representação oferece informações sobre a distribuição espectral ao longo do tempo, sendo valiosa para a análise de eventos temporais em sinais de áudio. Para facilitar a interpretação, muitas vezes aplica-se uma transformação logarítmica ao espectrograma, resultando no que é conhecido como espectrograma logarítmico.

No contexto de modelos multimodais, a representação de áudio através de espectrogramas permite a integração eficaz com dados visuais, promovendo a convergência de informações provenientes de diferentes modalidades sensoriais. A extração de características significativas a partir dessas representações é crucial para o sucesso da tarefa de aprendizado conjunto.

## 2.10 Sinais Auditivos

O sistema auditivo humano é uma das mais complexas e fascinantes partes do sistema sensorial, responsável por converter ondas sonoras em informações que podem ser proces-

Figura 2.7: Geração de um espectrograma



Fonte: Adaptado de Rankin (2017)

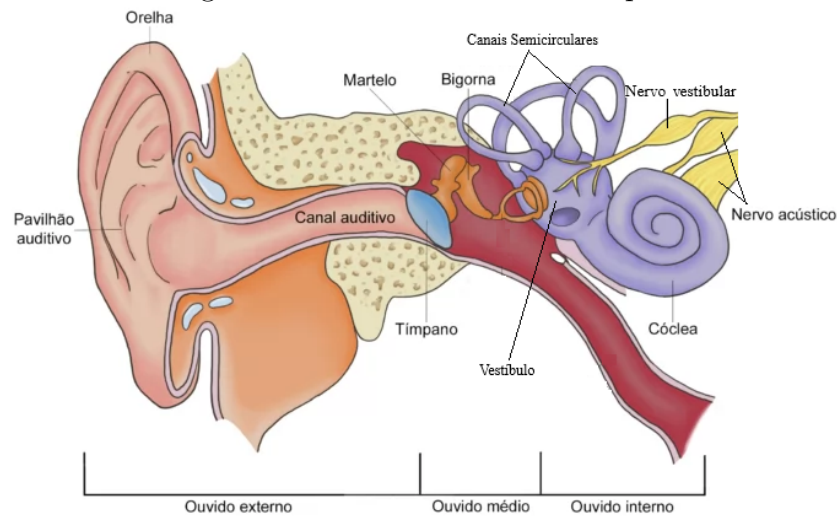
sadas pelo cérebro. Esta capacidade é essencial para a comunicação, a compreensão da fala e a percepção do ambiente.

### 2.10.1 Estrutura do Sistema Auditivo

O sistema auditivo receptor (externo), exposto na Figura 2.8, pode ser dividido em três partes principais: o ouvido externo, o ouvido médio e o ouvido interno:

- **Ouvido Externo:** Compreende a orelha externa, que inclui a parte visível da orelha e o canal auditivo externo. O pavilhão auricular ajuda a captar as ondas sonoras do ambiente e direcioná-las para o canal auditivo externo. O canal auditivo externo é revestido por uma camada de pele e glândulas que produzem cerume, o qual protege o canal e ajuda na manutenção de um ambiente adequado para a transmissão do som. No final do canal auditivo externo está o tímpano, uma membrana fina e delicada que vibra em resposta às ondas sonoras (MALMIERCA, 2015).
- **Ouvido Médio:** Consiste no espaço entre o tímpano e a cóclea, e é atravessado por três pequenos ossos chamados martelo, bigorna e estribo, também conhecidos como ossículos. Esses ossículos são responsáveis por amplificar as vibrações sonoras

Figura 2.8: Sistema auditivo receptor



Fonte: Adaptado de Barbosa (2018)

recebidas pelo tímpano e realizar sua transmissão para o ouvido interno (EHRET; ROMAND, 1997).

- **Ouvido Interno:** O ouvido interno, localizado após o ouvido médio, é formado pela cóclea, pelo vestíbulo e pelos canais semicirculares, além de se conectar ao cérebro por meio do nervo acústico e do nervo vestibular. A cóclea converte as vibrações sonoras em impulsos elétricos, que são transmitidos ao cérebro pelo nervo acústico. Já o vestíbulo e os canais semicirculares detectam movimentos e mudanças na posição da cabeça, contribuindo para o equilíbrio e a orientação espacial. As informações dessa região são conduzidas ao cérebro pelo nervo vestibular. Ambos os nervos se unem para formar o nervo vestibulococlear, responsável pela condução dos sinais de audição e equilíbrio (EHRET; ROMAND, 1997).

### 2.10.2 Processamento dos sinais elétricos

Os sinais elétricos gerados pelas células ciliadas da cóclea são enviados ao cérebro através do nervo auditivo. Esse processo envolve várias regiões do cérebro, começando pelo núcleo coclear no tronco cerebral, onde ocorre uma primeira análise dos sinais sonoros (ALBERTI, 2001).

Do tronco cerebral, os sinais seguem para o tálamo, uma estrutura cerebral crucial na transmissão de informações sensoriais para áreas específicas do córtex cerebral (MALMIERCA, 2015; EHRET; ROMAND, 1997). No caso da audição, o tálamo encaminha os sinais para o córtex auditivo primário, localizado no lobo temporal, e para outras

regiões associadas ao processamento auditivo, como o córtex auditivo secundário e áreas responsáveis pela integração multisensorial.

No córtex auditivo primário, ocorre a interpretação fundamental das características do som, como frequência, intensidade e duração. Essa região tem como principal função a identificação e discriminação de diferentes sons, permitindo ao cérebro reconhecer padrões sonoros e atribuir significado às informações auditivas recebidas (ALBERTI, 2001).

Além disso, o processamento auditivo no cérebro não se limita apenas à análise das propriedades básicas do som. Também inclui a capacidade de localizar a fonte sonora no espaço, discernir a direcionalidade do som, identificar a temporalidade dos eventos sonoros e integrar informações auditivas com outras modalidades sensoriais para uma percepção auditiva mais rica e contextualizada.

## 2.11 Sinais Visuais

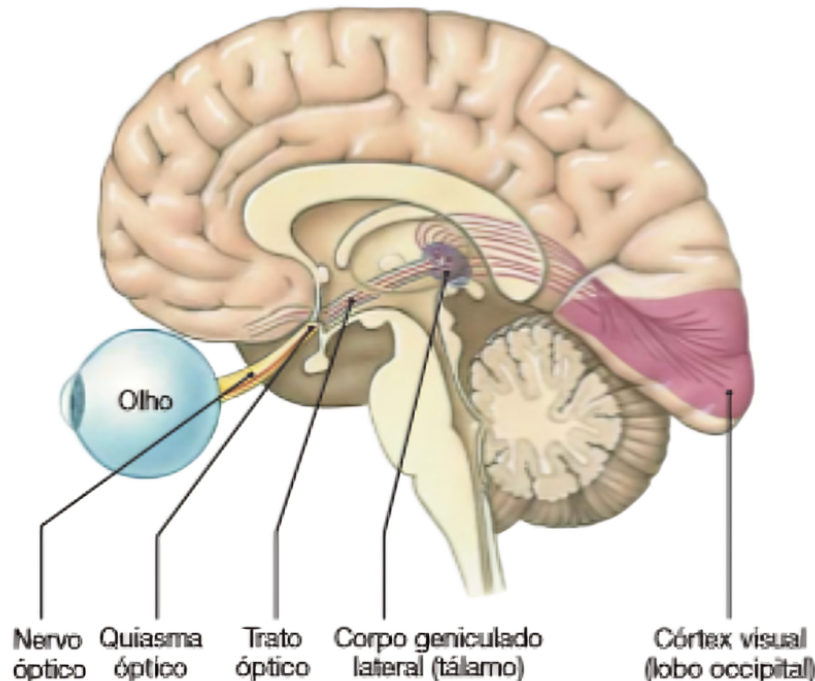
O sistema visual humano é fundamental para a percepção do mundo ao nosso redor, permitindo-nos detectar, interpretar e responder a estímulos visuais. Sendo de suma importância para a navegação, identificação de objetos e na comunicação não verbal.

### 2.11.1 Estrutura do Sistema Visual

O sistema visual humano, observado na Figura 2.9, é uma complexa rede de estruturas que trabalham em conjunto para captar, processar e interpretar informações visuais. Podemos olhar para a estrutura receptiva visual na seguinte estrutura:

- **Olhos:** Os olhos são os órgãos sensoriais primários da visão. Eles possuem várias estruturas essenciais para o processo visual, começando pela córnea, uma camada externa transparente que atua como uma lente inicial para direcionar a luz para dentro do olho. Logo após a córnea, o cristalino busca ajustar seu formato para focar a luz corretamente na retina (SCHILLER, 1986). Esse processo de ajuste do cristalino é chamado de acomodação e é fundamental para a focalização adequada da imagem na retina.
- **Retina:** Localizada na parte posterior do olho, a retina é uma camada de tecido que contém os fotorreceptores responsáveis por converter a luz em sinais elétricos. Esses fotorreceptores incluem os bastonetes, sensíveis à luz em condições de baixa luminosidade e responsáveis pela visão periférica e noturna, e os cones, que são responsáveis pela visão de cores e detalhes em condições de maior luminosidade (SCHILLER, 1986).

Figura 2.9: Sistema visual



Fonte: Adaptado de Brito (2017)

- **Nervo Óptico e Vias Visuais Centrais:** O nervo óptico é responsável por transmitir os sinais elétricos gerados pela retina até o cérebro. Esses sinais viajam pelas vias visuais centrais, que incluem o quiasma óptico, onde parte das fibras dos nervos ópticos se cruzam parcialmente (TOVÉE, 2008). Essa cruzamento parcial permite a integração de informações visuais de ambos os olhos em diferentes partes do cérebro. As vias visuais centrais incluem várias estruturas cerebrais, como o núcleo geniculado lateral do tálamo, que atua como um *hub* de processamento visual, e o córtex visual, onde ocorre a interpretação e a formação das imagens visuais.

### 2.11.2 Processamento dos sinais elétricos

Os sinais elétricos provenientes da retina viajam ao longo do nervo óptico até alcançar o quiasma óptico, uma estrutura crucial onde ocorre o parcial cruzamento das fibras nervosas, de modo que a informação visual de cada olho seja transmitida para os hemisférios cerebrais opostos. Esse cruzamento parcial permite uma integração eficiente das informações visuais recebidas pelos dois olhos (ARMSTRONG; CUBBIDGE, 2014).

Após o quiasma óptico, os sinais visuais são encaminhados ao núcleo geniculado lateral no tálamo, trabalhando como uma estação de retransmissão. O núcleo geniculado lateral não apenas repassa os sinais visuais para o córtex visual primário (V1), localizado no lobo

occipital do cérebro, mas também desempenha funções de processamento e filtragem das informações visuais antes de serem encaminhadas ao córtex visual primário.

O córtex visual primário, é uma região crucial para o processamento inicial das informações visuais (ARMSTRONG; CUBBIDGE, 2014). Nessa área do córtex, ocorre a interpretação básica das características visuais, como forma, cor, movimento e profundidade. Além disso, ele é de suma importância na organização e na representação topográfica do campo visual, permitindo ao cérebro criar uma representação coerente do mundo ao nosso redor.

## 2.12 Processamento Multimodal

A integração de informações auditivas e visuais no cérebro humano é um processo complexo e altamente coordenado que envolve várias regiões e mecanismos. Esse processamento multimodal é essencial para a percepção completa e precisa do ambiente, permitindo a detecção e interpretação de estímulos simultâneos provenientes de diferentes sentidos (SCHEEF et al., 2009).

### 2.12.1 Regiões Envolvidas no Processamento Multimodal

Duas das principais regiões envolvidas na integração auditiva e visual são o Sulco Temporal Superior (STS) e o Giro Temporal Superior (STG). Estas regiões recebem informações de ambos os córtex, e executam uma análise e interpretação de informações multisensoriais de forma síncrona (BEAUCHAMP et al., 2004).

O STS é conhecido por seu papel na integração de estímulos auditivos e visuais. Ele recebe e processa sinais de diferentes modalidades sensoriais, ajudando na criação de uma percepção coesa e unificada do ambiente. Estudos (SCHEEF et al., 2009; BESLE et al., 2008) mostram que o STS é ativado quando o cérebro está processando informações auditivas e visuais simultaneamente, facilitando a detecção de eventos multisensoriais e a coordenação de respostas apropriadas.

O STG, por sua vez, está envolvido na análise mais detalhada das características dos estímulos auditivos e visuais. Ele ajuda na decodificação das propriedades acústicas e visuais dos sinais, contribuindo para a identificação de fontes sonoras e visuais e para a compreensão de contextos mais complexos.



### 2.12.2 Fluxo de Informações Multimodal

O fluxo de informações multimodal no cérebro começa com a captação de estímulos sensoriais pelos órgãos receptores. Os olhos captam informações visuais, que são convertidas em sinais elétricos pela retina e transmitidas pelo nervo óptico para o córtex visual primário (SCHEEF et al., 2009; KNÖPFEL et al., 2019). Simultaneamente, as ondas sonoras são captadas pelo ouvido externo, amplificadas pelo ouvido médio e convertidas em sinais elétricos pela cóclea no ouvido interno. Esses sinais são enviados ao cérebro pelo nervo auditivo.

Após a captação inicial, os sinais auditivos e visuais seguem para o tálamo, que atua como um centro de retransmissão sensorial. No tálamo, ocorre uma primeira integração das informações sensoriais, e os sinais são direcionados para as áreas específicas do córtex cerebral: o córtex visual primário no lobo occipital e o córtex auditivo primário no lobo temporal.

O processamento detalhado ocorre nas áreas associativas, onde o STS e o STG são de suma importância. O STS integra as informações auditivas e visuais, ajudando na percepção de eventos multisensoriais como a sincronização de sons com movimentos visuais (BEAUCHAMP et al., 2004). Já o STG analisa e refina as características dos estímulos, permitindo a identificação precisa e a interpretação dos sinais auditivos e visuais.

Pesquisas iniciais sugeriam que a integração multimodal ocorria em um processo hierárquico, apenas após o processamento inicial nos córtex individuais. No entanto, estudos mais recentes indicam que o processamento das informações auditivas e visuais nos córtices primários pode ser influenciado de maneira intermodal durante os estágios iniciais. Nosso objetivo é explorar o processamento homogêneo e a correlação de dados audiovisuais, inspirado nos processos do cérebro, para reproduzir a integração sensorial que caracteriza a representação multimodal biológica. (BEAUCHAMP et al., 2004). Essas descobertas apontam para uma interação mais dinâmica entre os diferentes tipos de estímulos sensoriais, promovendo uma percepção integrada e adaptativa do ambiente.

Além dessas áreas, o *Putâmen* também contribui para o processamento multimodal, especialmente na coordenação temporal e na sincronização dos estímulos (SCHEEF et al., 2009). O *Putâmen* ajuda a alinhar as informações auditivas e visuais no tempo, garantindo que o cérebro perceba eventos simultâneos de maneira coesa. Esse alinhamento temporal é fundamental para a integração precisa dos estímulos, como observado em fenômenos audiovisuais, como o efeito *McGurk* (MCGURK; MACDONALD, 1976), onde a percepção do som é alterada pela presença de estímulos visuais conflitantes, demonstrando a complexidade e a interdependência do processamento multimodal no cérebro humano.

### **2.12.3 Integração e Percepção**

A integração das informações auditivas e visuais no cérebro resulta em uma percepção enriquecida e contextualizada. Esta integração permite que o cérebro combine características de diferentes modalidades sensoriais para formar uma representação unificada do ambiente (MCGURK; MACDONALD, 1976; BEAUCHAMP et al., 2004). Por exemplo, a sincronização de um som com um movimento visual ajuda na localização da fonte sonora e na compreensão do evento observado.

O processamento multimodal é, portanto, uma característica essencial do sistema sensorial humano, permitindo a navegação eficiente e a interação com o ambiente de maneira precisa e adaptativa. Estudar e emular esses processos no desenvolvimento de modelos computacionais pode levar a avanços significativos em sistemas de percepção artificial, proporcionando uma compreensão mais profunda e uma interação mais natural com o mundo ao nosso redor.

## 3 Trabalhos Relacionados

A revisão dos trabalhos relacionados desempenha um papel crucial na fundamentação e na construção do contexto para esta pesquisa, que visa a convergência multimodal entre áudio e imagem. Nosso olhar crítico volta-se para estudos que transcendem as barreiras tradicionais entre diferentes modalidades sensoriais, especialmente na convergência entre áudio e imagem. Inspiramo-nos nessas contribuições para enriquecer e aprimorar nossa abordagem, buscando compreender as nuances e os desafios enfrentados por pesquisadores que exploraram a união de diversas modalidades, sejam elas visuais, linguísticas ou auditivas.

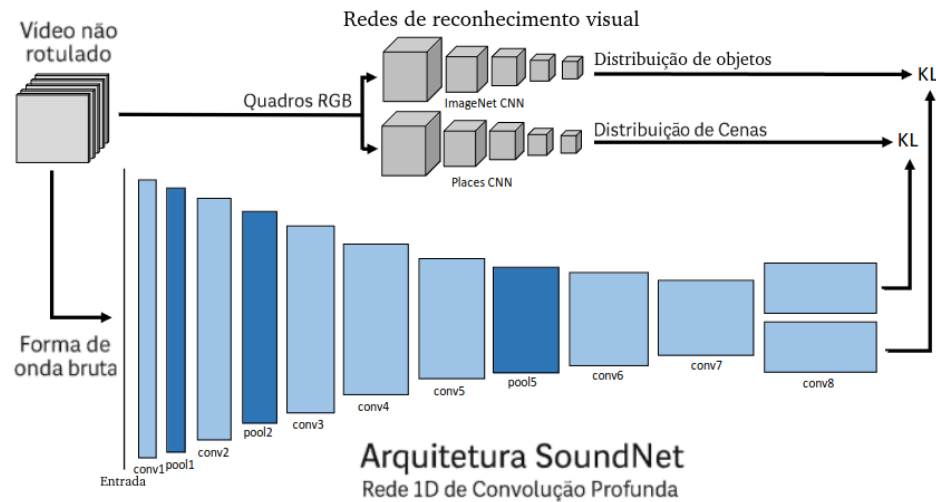
### 3.1 Trabalhos do Meio da Computação

Em seu estudo Noda et al. (2015), exploraram a associação entre movimentos labiais visuais e palavras faladas, demonstrando o potencial das pistas visuais para o reconhecimento da fala. O artigo trabalha com as duas modalidades ao empregar duas redes específicas para extração de características de ambas as modalidades. Uma terceira rede é então encarregada de processar as representações geradas para as duas modalidades e realizar o reconhecimento da fala. Esse estudo demonstra a eficácia de aproveitar o movimento dos lábios como uma modalidade adicional para melhorar a precisão do reconhecimento da fala.

Buscando fazer o uso da grande quantidade de vídeos não catalogados, Aytar, Vondrick e Torralba (2016) mergulharam no domínio da aprendizagem de representações sonoras apenas a partir de dados de vídeo, sem supervisão explícita de áudio. Ao treinar redes neurais convolucionais profundas em conjuntos de dados de vídeo em grande escala, como exibido na Figura 3.1, revelaram que a rede dedicada à extração de representações auditivas pode ser treinada exclusivamente ao examinar as representações de modelos visuais relacionadas à porção visual do mesmo trecho de vídeo. Este estudo sugere que as representações visuais e auditivas extraídas por redes neurais estão inter-relacionadas e podem complementar-se mutuamente no processamento de dados audiovisuais.

Também buscando fazer o uso da grande quantidade de dados não catalogados, Arandjelovic e Zisserman (2017) introduziram uma abordagem para a aprendizagem de represen-

Figura 3.1: Processo de treinamento proposto para treinamento da Soundnet



Fonte: Adaptado de Aytar, Vondrick e Torralba (2016)

tações multimodais através do treino de redes neurais profundas num conjunto de dados em grande escala constituído por dados de áudio e imagens emparelhadas. O processo de treinamento foi inicialmente concebido como um problema de classificação para verificar se o trecho do áudio corresponde a um determinado frame do vídeo. Para alcançar isso, foi proposto o uso e treinamento conjunto de duas sub-redes dedicadas à extração de características de cada modalidade, realizando o processamento de forma separada para os dados de áudio e vídeo. Cada sub-rede demonstrou a capacidade de gerar representações coerentes para sua modalidade específica, alcançando desempenho satisfatório em tarefas de classificação atribuídas a cada modalidade. Novamente podemos observar o efeito positivo de utilizar informações de ambas as modalidades para o processamento de dados audiovisuais

Podemos também buscar inspiração da junção de outras modalidades. No âmbito de processamento de documentos ricos visualmente, o processamento conjunto de informações visuais e textuais pode ser usado de forma benéfica para solução de problemas complexos. Huang et al. (2022), por exemplo, propõem a aplicação de técnicas de treinamento autossupervisionado para realizar o pré-treinamento de uma rede transformers, visando extrair informações valiosas de documentos diversos. Ao mascarar aleatoriamente trechos das informações textuais e visuais, foi possível treinar sobre um grande conjunto de dados não catalogados. O modelo é então capaz de processar e gerar representações vetoriais (*embedding*) de um documento que podem então ser usadas em tarefas como classificação de documentos e extração de tabelas em documentos estruturados. Ao adotar uma abordagem que considera tanto a representação visual quanto a linguística, o estudo alcança resultados notáveis na eficiência do entendimento multimodal.

Diferente das abordagens que exploramos até aqui, Girdhar et al. (2023b) buscam a representação em um mesmo espaço para diversas modalidades. Temos a representação conjunta para dados textuais, visuais, auditivos, termais, de profundidade e dados egocêntricos . Ao explorar a criação de um espaço de características unificado, propõe uma solução para a convergência multimodal que usa como partida o treinamento de redes codificadoras de cada uma das modalidades pareadas. Todas as redes são treinadas para gerar uma codificação compatível com as representações geradas pelas redes de outras modalidades para dados de um mesmo elemento. A pesquisa abrange efetivamente a inter-relação entre modalidades distintas, proporcionando uma representação compartilhada que transcende as barreiras tradicionais entre as modalidades do estudo.

Com relação direta ao nosso ponto de partida, buscamos também base em outros trabalhos que exploraram a relação entre áudio e imagem.

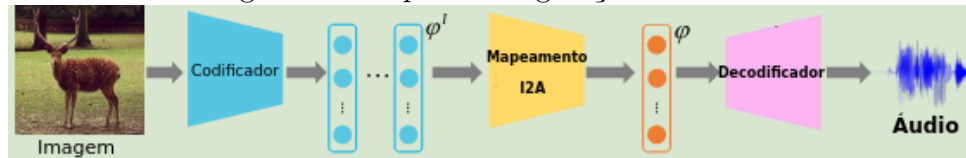
Żelaszczyk e Mańdziuk (2022) de forma similar ao nosso projeto, buscaram investigar a síntese de imagens a partir de dados de áudio utilizando Autoencoders Variacionais . O experimento é fundamentado no conjunto de dados do Modified National Institute of Standards and Technology (MNIST) e em gravações relacionadas aos números representados verbalmente. Diversas combinações de exemplos foram testadas em termos de paridade. No treinamento da rede, uma segunda rede discriminadora, preexistente e previamente treinada, foi empregada para classificar as imagens geradas, fornecendo uma métrica de desempenho para o modelo. Os resultados revelaram que a rede generativa alcançou sucesso na geração de imagens realistas, com a rede discriminadora atingindo uma acurácia de 94,2% na classificação das imagens geradas.

Yang, Su e Wang (2020) também propuseram a geração de imagens a partir de dados de áudio, empregando a SoundNet como codificador das características dos áudios e adotando a técnica de Conditional Generative Adversarial Network (cGAN) para a geração de imagens . O conjunto de dados compreende 10.701 pares de áudio e imagens, abrangendo diversas situações. O modelo treinado demonstrou um sucesso relativo, alcançando uma acurácia de 86% quando avaliado por meio de um classificador nas imagens geradas.

No caminho contrário dos artigos anteriores, Ning et al. (2021) exploraram a geração de áudio para descrever imagens variadas . Propõe uma abordagem de geração não direta entre imagem e áudio, onde inicialmente a imagem é codificada em uma transformação vetorial (*embedding*). A partir dessa *embedding*, ocorre o mapeamento para uma representação correspondente do áudio (também *embedding*), e somente então essa representação é decodificada para produzir o áudio final, completando o processo de geração. O processo completo pode ser observado na Figura 3.2. O treinamento da rede foi conduzido utilizando um conjunto diversificado de dados, englobando conjuntos de dados renomados, como o MNIST, e dois conjuntos do Canadian Institute For Advanced Research (CIFAR), nomeadamente CIFAR100 e CIFAR10, os quais representam subconjuntos com 100 e

10 classes, respectivamente. Essa abordagem alcançou sucesso relativo, demonstrando desempenho comparável ao estado da arte nesse domínio.

Figura 3.2: Pipeline de geração de áudio



Fonte: Adaptado de Ning et al. (2021)

Analisando trabalhos que exploram o processamento conjunto e sinérgico das modalidades visual e auditiva, é possível observar os esforços para aprofundar a compreensão do comportamento dos modelos atuais em relação a esse problema.

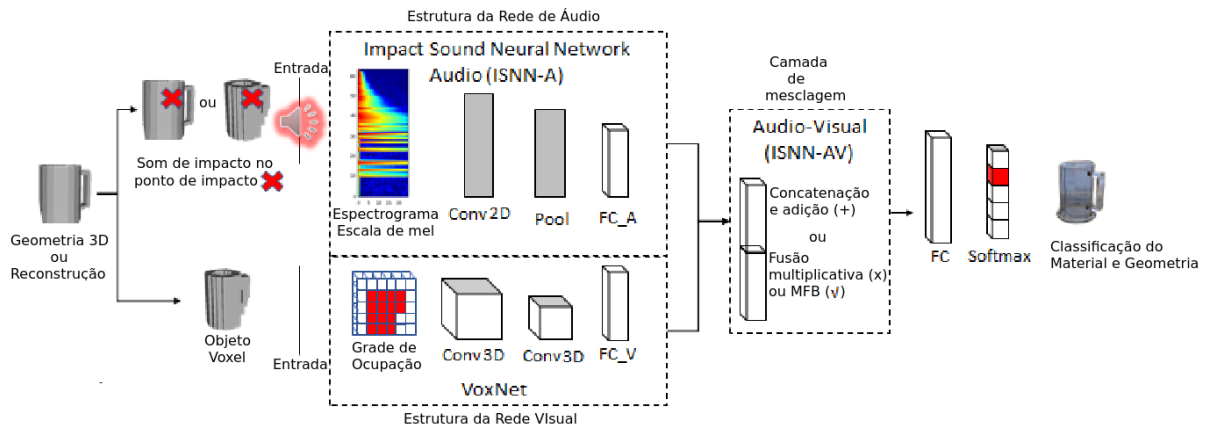
Lin e Wang (2020) propõem uma estrutura de aprendizado profundo utilizando um mecanismo de coatenção entre modalidades para localizar eventos em vídeos, analisando conjuntamente informações visuais e de áudio. O modelo, denominado AV-transformer, aproveita dados visuais intra e entre quadros, junto com características de áudio, para realizar coatenção entre essas modalidades. Essa abordagem melhora a extração de características espaciais e temporais informativas, melhorando o desempenho na localização de eventos. O modelo também introduz atenção a nível de instância, permitindo a identificação de regiões específicas da imagem associadas aos sons ou eventos relevantes. Resultados experimentais em conjuntos de dados de referência validam a eficácia do AV-transformer, demonstrando sua capacidade em tarefas de localização de eventos audiovisuais.

Já Sterling et al. (2018) apresentam uma rede neural multimodal otimizada para estimar a geometria e o material de objetos utilizando sons de impacto. A Impact Sound Neural Network - Áudio (ISNN-A) usa espectrogramas de sons de objetos reais ou sintéticos, enquanto a rede ISNN-AV, observada na Figura 3.3, combina ISNN-A com VoxNet para classificação de geometria e material de objetos. A avaliação em múltiplos conjuntos de dados demonstra a eficácia da rede em reconstruções de cena em tempo real, permitindo a classificação e segmentação de objetos transparentes ou ocultos visualmente. O método alcançou resultados promissores, superando abordagens baseadas apenas em visão em cenários desafiadores.

## 3.2 Trabalhos do Meio Biológico

Também com base na nossa inspiração, temos alguns trabalhos que exploram a relação de processamento auditivo e visual no cérebro humano, servindo como uma inspiração

Figura 3.3: Impact Sound Neural Network - AV



Fonte: Adaptado de Sterling et al. (2018)

sólida para a nossa pesquisa

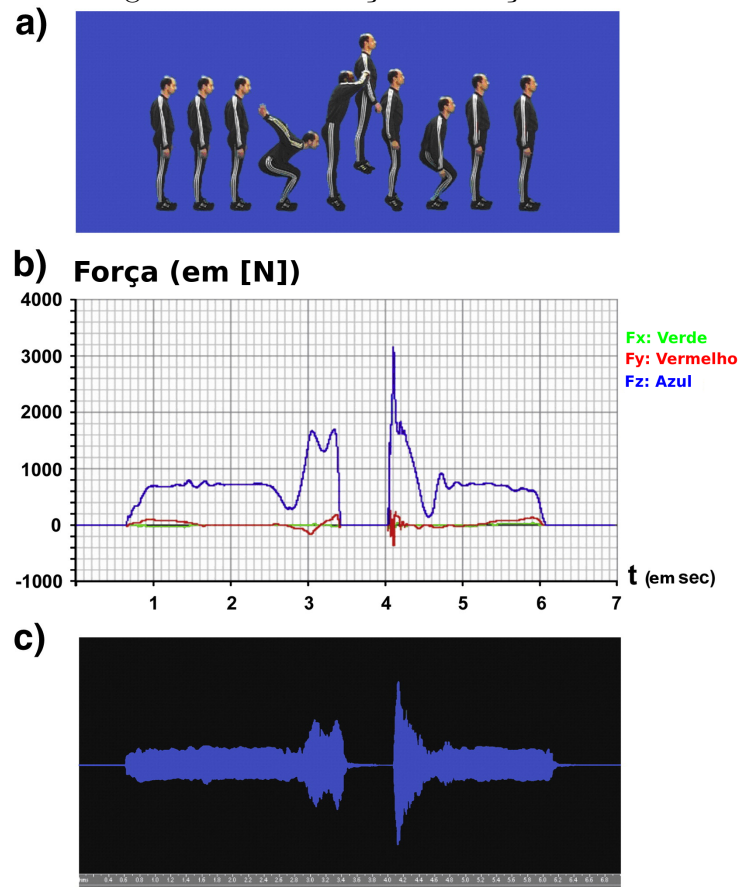
O trabalho de Knöpfel et al. (2019) aborda o impacto da experiência audiovisual na formação de conjuntos multissensoriais no córtex visual de camundongos adultos. O estudo investiga como a exposição conjunta a estímulos auditivos e visuais influencia a organização neuronal no córtex visual. Os resultados indicam que a experiência audiovisual fortalece a formação de conjuntos neuronais que respondem a estímulos tanto auditivos quanto visuais. Essa descoberta sugere uma plasticidade significativa no córtex visual em resposta a estímulos multissensoriais, destacando a adaptabilidade do sistema nervoso a experiências audiovisuais e seu papel na configuração das respostas neuronais no contexto multimodal.

Explorando a mesma relação entre os sentidos auditivos e visuais, Besle et al. (2008) exploram as interações entre a percepção visual e auditiva durante a compreensão da fala. Utilizando gravações intracranianas em seres humanos, o estudo investiga como a ativação visual influencia a atividade cortical auditiva durante a percepção da fala. Os resultados sugerem uma interconexão complexa entre os sistemas visual e auditivo no córtex auditivo, destacando a importância das interações multimodais na compreensão da linguagem. Este trabalho oferece dicas valiosas para o entendimento dos processos neurais subjacentes à integração de informações visuais e auditivas durante a percepção da fala humana.

Scheef et al. (2009) exploram como o cérebro humano integra informações audiovisuais para melhorar a percepção e as tarefas motoras. Utilizando como base a percepção de movimentação e utilizando a técnica de sonificação para transformar dados de força de reação do solo em sinais auditivos (observado na Figura 3.4), o estudo apresentou estímulos áudio-visuais concordantes e discordantes durante escaneamentos de fMRI. Os resultados revelaram aumento da ativação na área V5/MT para estímulos concordantes, indicando

seu papel na integração multimodal de movimentos, além do processamento visual.

Figura 3.4: Sonificação da reação do solo



Fonte: Adaptado de Scheef et al. (2009)

Besle et al. (2008) investigam como o córtex auditivo humano responde a estímulos visuais e auditivos durante a percepção da fala, utilizando registros intracranianos em pacientes epiléticos. Os resultados mostram que o córtex auditivo é ativado por movimentos labiais visuais, com interações audiovisuais ocorrendo rapidamente após o início do som. Estas descobertas fornecem informações valiosas sobre a cronologia e os mecanismos das interações multimodais no cérebro humano.

Garner e Keller (2022) em seu artigo na Nature Neuroscience investigam a integração de sinais auditivos e visuais no cérebro de um rato utilizando técnicas avançadas de neuroimagem, eletrofisiologia e modelagem computacional. O estudo destaca como o sulco temporal superior (STS) e o giro temporal superior (STG) são de extrema importância na fusão dessas modalidades sensoriais. Os resultados indicam que a coordenação temporal e a sincronização dos estímulos são significativamente influenciadas pelo putâmen. Além disso, os autores descobriram que a integração multimodal influencia os córtices auditivo e visual primários desde os estágios iniciais de processamento, desafiando a visão hierárquica tradicional e sugerindo um modelo interconectado de processamento sensorial.



Já Hasegawa et al. (2004) exploraram como associações aprendidas entre estímulos auditivos e visuais são processadas no cérebro. Utilizando fMRI, os pesquisadores investigaram a ativação do planum temporale (PT) em pianistas bem treinados ao observarem movimentos de teclas de piano. A pesquisa envolveu três grupos: indivíduos sem experiência em piano, com pouca experiência e bem treinados. Os resultados mostraram que apenas os pianistas bem treinados apresentaram ativação significativa no PT esquerdo ao observar movimentos de teclas, independentemente da familiaridade com a música. Isso sugere que o PT está envolvido em mapear sequências visuais complexas para sequências auditivas, refletindo um processo aprendido que ocorre antes da identificação musical.

Estes trabalhos relacionados realçam coletivamente a importância da aprendizagem multimodal e da fusão das modalidades áudio, visual e outras. Mostram o potencial da utilização de diversas fontes de dados para melhorar o desempenho e diminuir a lacuna entre diferentes domínios sensoriais.

### 3.3 Classificação de Vídeos

Modelos com capacidade de processamento multimodal possuem ampla aplicação em tarefas audiovisuais, abrangendo desde a legendagem de vídeos e ferramentas de acessibilidade até soluções voltadas ao setor de entretenimento. Com o objetivo de validar a arquitetura proposta, esta pesquisa se restringirá à tarefa de classificação de vídeos, o que permite um foco mais específico e uma validação inicial consistente do modelo desenvolvido.

A tarefa de classificação de vídeos é uma das aplicações mais tradicionais da visão computacional. A imensa quantidade de conteúdo audiovisual disponível torna essencial o desenvolvimento de ferramentas cada vez mais eficientes para a classificação e organização dessa mídia. Desde os primórdios da área, inúmeros trabalhos têm sido desenvolvidos com avanços significativos nesse domínio. Nos parágrafos a seguir, exploraremos algumas dessas pesquisas, com o intuito de estabelecer uma base comparativa para a avaliação da nossa abordagem.

Um dos conjuntos de dados mais utilizados para experimentos e validação de abordagens em classificação de vídeo é o *UCF101*. Esse conjunto contém vídeos curtos, organizados em 101 categorias de ações humanas, sendo amplamente adotado para avaliação de modelos que exploram tanto informações visuais quanto temporais. A complexidade da tarefa decorre não apenas da variabilidade visual entre os vídeos, mas também da necessidade de capturar relações temporais entre *frames* consecutivos. Redes recorrentes (*RNNs*) foram amplamente utilizadas nesse contexto para modelar a dinâmica temporal, entretanto sua eficiência é limitada em sequências longas. A ausência de um mecanismo robusto de

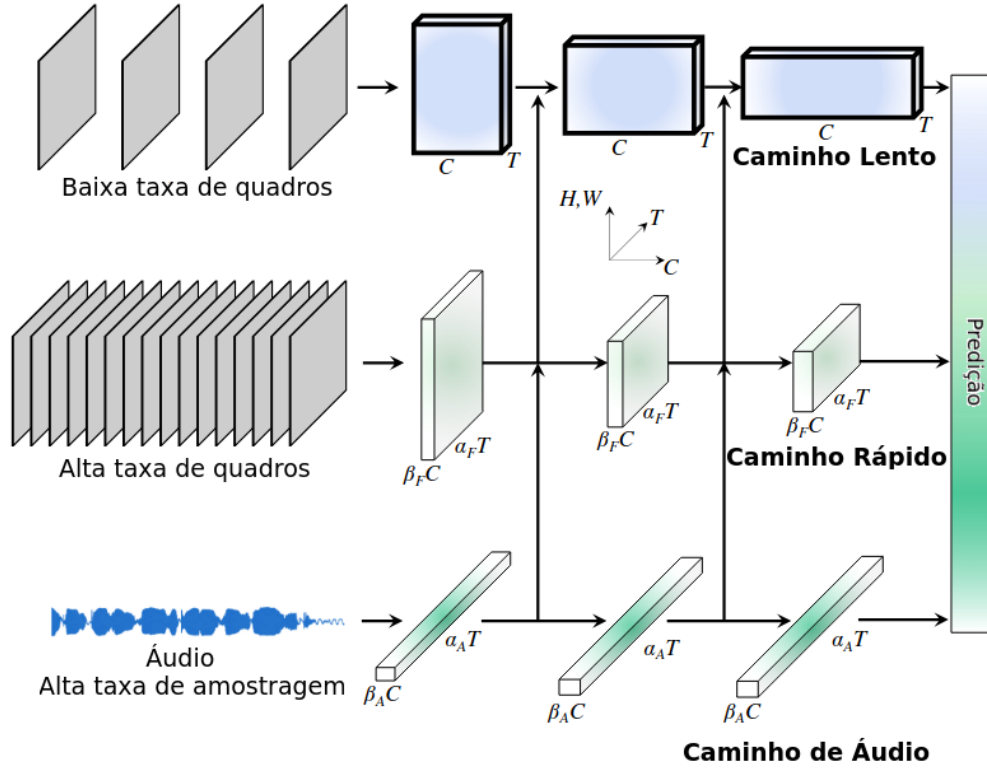
memória compromete a capacidade dessas redes de operar de forma eficaz nesses cenários. A seguir, revisamos algumas das abordagens mais relevantes nessa tarefa, com ênfase nos conjuntos *UCF101* e *COIN*.

Alfasly et al. (2024) propõem uma arquitetura multimodal inovadora voltada para reconhecimento de ações em vídeos anotados exclusivamente com informações visuais, como o UCF-101. O método emprega múltiplas modalidades — visual (RGB e fluxo óptico), áudio e textual — organizadas em um sistema de fusão adaptativa. O modelo utiliza uma estratégia chamada *Irrelevant Modality Dropout* (IMD), que avalia a relevância entre modalidades por meio de uma rede neural treinável, descartando modalidades irrelevantes para cada amostra individual. A ligação semântica entre os rótulos visuais e auditivos é estabelecida por meio de dicionários gerados com modelos de linguagem como BERT e GPT-3. A arquitetura também inclui uma etapa de reordenação das previsões com auxílio do GPT-3, utilizando objetos visuais detectados e sons estimados. Avaliada no conjunto UCF-101, a abordagem alcança uma acurácia Top-1 de 99,3%, superando métodos multimodais anteriores e posicionando-se como o estado da arte atual para essa tarefa.

Xiao et al. (2020) propõem o modelo *AVSlowFast*, uma extensão multimodal do *SlowFast* voltada para a fusão audiovisual hierárquica. A arquitetura, exibida na Figura 3.5, introduz um terceiro caminho dedicado ao áudio, baseado em espectrogramas log-mel, que opera em alta resolução temporal e espectral, sendo integrado às camadas intermediárias da rede visual. Um mecanismo chamado *DropPathway* é proposto para regularizar o treinamento ao desativar aleatoriamente o caminho auditivo, promovendo maior robustez na presença de ruído ou ausência de sinal. Além disso, os autores incorporam tarefas de sincronização audiovisual como forma de aprendizado autossupervisionado auxiliar. Avaliado no conjunto UCF101, o modelo *AVSlowFast* alcançou uma acurácia Top-1 de 87,0%, superando variantes multimodais e destacando a eficácia da integração audiovisual mesmo em tarefas supervisionadas tradicionais. A principal diferença em relação à nossa proposta é que o *AVSlowFast* realiza a fusão entre áudio e vídeo de forma hierárquica e precoce, injetando características auditivas diretamente nas camadas intermediárias do extrator visual, enquanto a X-Net opta por processar as modalidades de forma independente e realizar a fusão por modulação condicional (via FiLM) após a extração unimodal. Essa escolha torna a X-Net mais modular e facilmente extensível para outras modalidades, ao custo de não explorar a fusão precoce explorada pelo *AVSlowFast*. Na Figura 3.5, observamos três caminhos de processamento: o **Caminho Lento** (topo), que processa o vídeo em baixa taxa de quadros ( $T$ ) priorizando contexto semântico; o **Caminho Rápido** (meio), que opera em alta taxa de quadros ( $\alpha_F T$ ) para capturar detalhes de movimento, com redução de canais ( $\beta_F C$ ); e o **Caminho de Áudio** (base), que processa espectrogramas de áudio em alta taxa de amostragem ( $\alpha_A T$ ), também com redução de canais ( $\beta_A C$ ). As setas indicam conexões laterais entre caminhos, permitindo a troca de

informações multimodais antes da previsão final. As variáveis  $C$ ,  $T$ ,  $\alpha$  e  $\beta$  representam, respectivamente, número de canais, dimensão temporal, fator de ampliação temporal e fator de redução de canais. Esse design garante que cada fluxo opere com granularidade temporal distinta, mas compartilhe representações no momento da fusão.

Figura 3.5: Modelo AvSlowFast

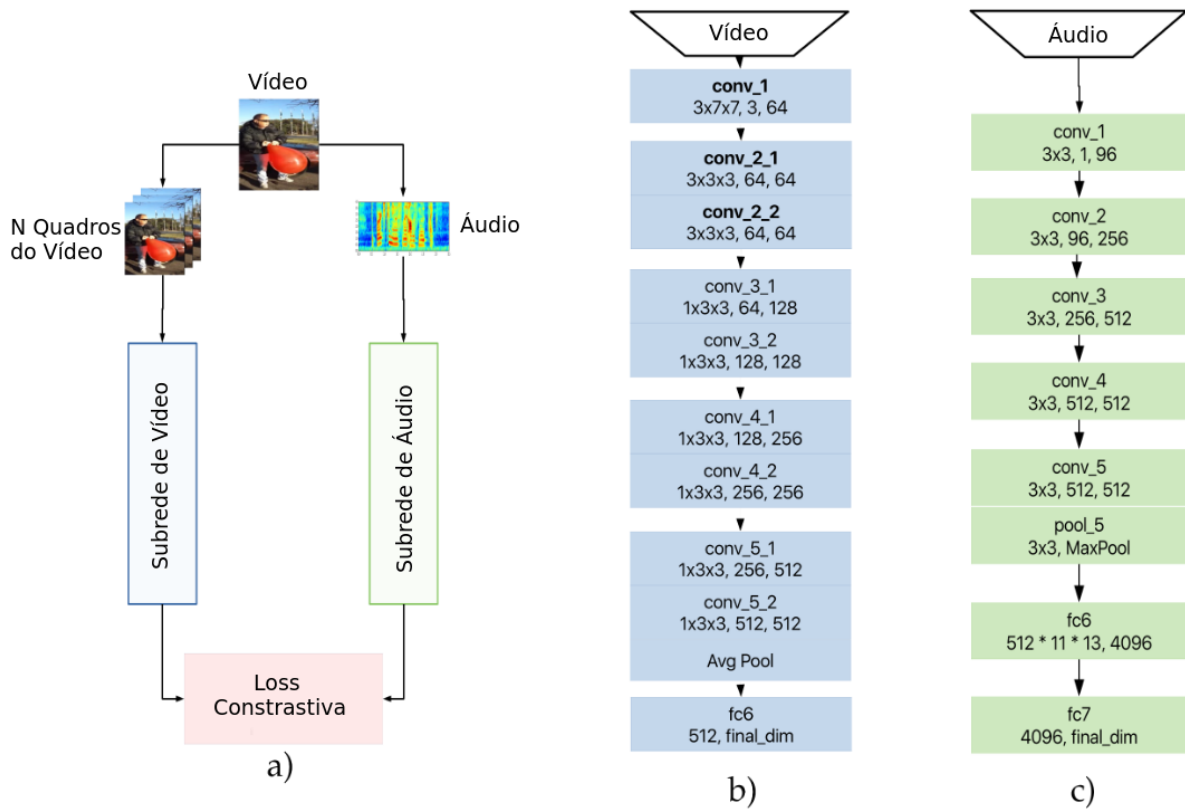


Fonte: Adaptado de Xiao et al. (2020)

Korbar, Tran e Torresani (2018) propõem um modelo autossupervisionado baseado na tarefa de sincronização temporal entre áudio e vídeo, denominada *Audio-Visual Temporal Synchronization* (AVTS). A arquitetura, observada na Figura 3.6, é composta por duas *streams*, uma para vídeo e outra para áudio, que são treinadas conjuntamente com perda contrastiva para distinguir pares sincronizados de pares dessincronizados. O diferencial do trabalho está na introdução de *negativos difíceis*, gerados a partir de pares extraídos de diferentes momentos do mesmo vídeo, e no uso de *curriculum learning* para facilitar a convergência. Após o pré-treinamento autossupervisionado no dataset Kinetics, o modelo é ajustado para a tarefa de reconhecimento de ações. No conjunto de dados UCF101, o modelo baseado na arquitetura MC3 alcança uma acurácia de 85,8% quando pré-treinado com AVTS. A principal diferença em relação à nossa proposta é que o AVTS utiliza a tarefa de sincronização apenas como um estágio de pré-treinamento para melhorar a qualidade das representações, enquanto a X-Net incorpora o módulo SyncAV como parte ativa e funcional no pipeline de classificação. Na X-Net, o SyncAV é treinado separadamente

para detectar alinhamento temporal e suas representações são fundidas com reconstruções cruzadas geradas pelo ProcessorAV, produzindo *embeddings* multimodais interpretáveis que influenciam diretamente as predições finais. Em contraste, o AVTS não explora explicitamente o uso contínuo das representações de sincronização durante a inferência, mas apenas como supervisão indireta no aprendizado inicial. Na Figura 3.6, o diagrama (a) mostra a divisão do processamento em duas sub-redes: uma sub-rede de vídeo (blocos azuis) e uma sub-rede de áudio (blocos verdes). O vídeo é convertido em  $N$  quadros e processado por convoluções 3D (vide **b**), enquanto o áudio é convertido em bancos de filtros espectrais e processado por convoluções 2D (vide **c**). As dimensões especificadas, como  $3 \times 7 \times 7, 3, 64$ , indicam respectivamente tamanho do kernel, número de canais de entrada e número de filtros de saída. As saídas de ambas as sub-redes são projetadas para vetores de dimensão fixa ( $fc6$ ) e comparadas por perda contrastiva para identificar pares sincronizados e dessincronizados.

Figura 3.6: Modelo AVTS



Fonte: Adaptado de Korbar, Tran e Torresani (2018)

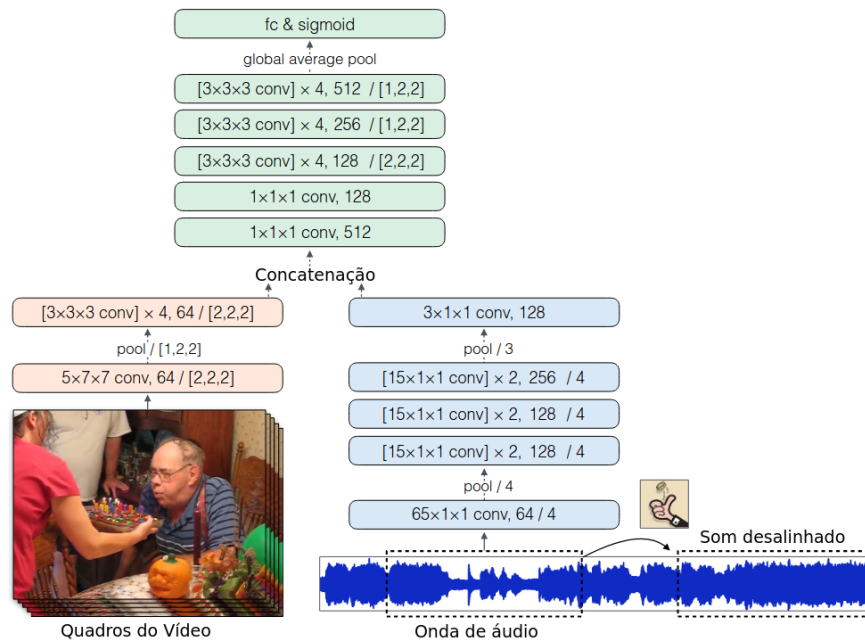
Huang et al. (2018) propõem uma análise crítica sobre o papel da informação temporal em modelos de compreensão de vídeo, questionando até que ponto o movimento é realmente essencial para a tarefa de reconhecimento de ações. Para isso, introduzem dois mecanismos analíticos: um gerador temporal agnóstico de classe e um seletor de quadros invariante ao

movimento, que permitem ablações sistemáticas da informação temporal sem modificar os pesos do modelo. Utilizando o modelo C3D como base, os autores mostram que é possível reduzir a lacuna de desempenho causada pela remoção do movimento de 25% para apenas 6% no dataset UCF101, demonstrando que muitos modelos atuais não exploram adequadamente a informação temporal disponível. Mesmo com apenas um quadro por segmento, o modelo ainda atinge desempenho competitivo, e com o vídeo completo, atinge uma acurácia de 82,3% no UCF101. Essa descoberta é particularmente relevante para arquiteturas que buscam eficiência computacional, uma vez que a retenção de apenas 4 quadros pode alcançar resultados comparáveis ao uso de toda a sequência. O trabalho fornece uma perspectiva valiosa sobre as limitações dos modelos atuais e destaca a importância da análise qualitativa e quantitativa da influência do movimento em tarefas de classificação de vídeos.

Owens e Efros (2018) propõem um modelo autossupervisionado para análise multissensorial de cenas, treinado com base na detecção de sincronização temporal entre fluxos de áudio e vídeo. O modelo, ilustrado na Figura 3.7, utiliza uma arquitetura convolucional 3D de fusão precoce (*early fusion*), onde sequências de *frames* de vídeo e formas de onda de áudio passam por caminhos de convolução independentes, sendo posteriormente concatenados e processados conjuntamente para prever se os fluxos estão ou não alinhados no tempo. A representação aprendida é então aplicada a diversas tarefas, como reconhecimento de ações, localização de fontes sonoras e separação de áudio *on/off-screen*. Para o reconhecimento de ações, o modelo é pré-treinado na tarefa de sincronização usando vídeos do AudioSet e posteriormente ajustado no dataset UCF101. Os resultados mostram que, mesmo sem rótulos supervisionados, o modelo atinge uma acurácia de 82,1% no UCF101, superando diversos métodos autossupervisionados anteriores, como Shuffle-and-Learn (50,9%) e O3N (60,3%). Apesar de também explorar a tarefa de sincronização, a principal diferença em relação à nossa proposta é que Owens e Efros utilizam o alinhamento como objetivo único de pré-treinamento, visando gerar representações transferíveis para outras tarefas. Em contraste, na X-Net o módulo SyncAV é treinado separadamente e suas representações de sincronização são combinadas com reconstruções cruzadas no módulo de saída multimodal, modulando diretamente as *backbones* durante a classificação. Assim, a sincronização não é apenas um pré-processo, mas um componente funcional integrado ao fluxo de decisão final.

Xu et al. (2019) propuseram uma abordagem autosupervisionada para aprendizado espaciotemporal em vídeos, baseada na predição da ordem de segmentos temporais (*clip order prediction*). Em contraste com métodos anteriores que utilizam apenas quadros individuais, a técnica explora diretamente a dinâmica dos vídeos ao embaralhar e ordenar pequenos cliques, utilizando arquiteturas 3D convolucionais como C3D, R3D e R(2+1)D. Após o pré-treinamento autossupervisionado, os modelos são refinados para a tarefa de reconhecimento de ações, resultando em melhorias significativas sobre inicializações

Figura 3.7: Estrutura Multisensory.



Fonte: Adaptado de Owens e Efros (2018)

aleatórias. Na base UCF101, a versão R(2+1)D do modelo atinge uma acurácia de 72,4%, destacando-se entre os métodos autossupervisionados sem pré-treinamento supervisionado. A abordagem mostra-se promissora para o aprendizado de representações multimodais robustas, especialmente em cenários com escassez de dados anotados.

Com o objetivo de lidar com vídeos longos em tarefas multimodais, He et al. (2024) propõem o modelo *Memory-Augmented Large Multimodal Model* (MA-LMM). A principal inovação consiste no uso de um banco de memória de longo prazo que armazena informações visuais processadas de forma sequencial, evitando a limitação de contexto dos *Large Language Models* (LLMs) e reduzindo o consumo de memória da GPU. Ao contrário de abordagens anteriores, como *Video-LLaMA* e *VideoChatGPT*, que processam todos os quadros simultaneamente, o MA-LMM opera de maneira autoregressiva, incorporando a cada passo as representações extraídas por um *visual encoder* pré-treinado e refinadas por um módulo de alinhamento baseado em *Querying Transformer*. O banco de memória armazena tanto *features* visuais brutas quanto *queries* aprendidas, que são comprimidas usando uma técnica baseada em similaridade temporal. Essa estrutura, permite ao modelo recuperar e contextualizar informações passadas de forma eficiente. Avaliado em tarefas de classificação de atividades em vídeos longos, o MA-LMM superou os modelos anteriores ao atingir 93,0% de acurácia no conjunto *Breakfast* e 93,2% no *COIN*, destacando-se como o novo estado da arte nessas tarefas.

Li et al. (2024) investigam a aplicação de modelos do tipo *state-space*, especificamente a

arquitetura *Mamba*, para tarefas de classificação de vídeos. O trabalho propõe a substituição de camadas temporais tradicionais em modelos de vídeo por blocos *Mamba*, que são mais eficientes em termos de memória e escalabilidade. A arquitetura final, denominada *Video Mamba*, é avaliada em diversos cenários, destacando-se pelo seu desempenho competitivo. No conjunto de dados COIN, o modelo alcança uma acurácia Top-1 de 90,5%, superando diversos métodos baseados em Transformers.

Islam et al. (2023) propõem o modelo *TranS4mer*, uma arquitetura eficiente para detecção de cenas em vídeos longos, como filmes, que combina mecanismos de atenção e modelos de estado espaço estruturado. A arquitetura é baseada em blocos *S4A*, que integram *self-attention* para capturar dependências de curto alcance entre quadros de uma mesma tomada, e operadores *Structured State Space* (S4) com portas de ativação (*Gated S4*) para modelar relações de longo prazo entre tomadas distintas. O modelo opera diretamente sobre sequências de *patches* extraídos dos quadros e é treinado de forma autosupervisionada com estratégias contrastivas, sendo posteriormente ajustado com anotação supervisionada. Avaliado em tarefas de classificação de atividades em vídeos longos, o *TranS4mer* superou abordagens anteriores baseadas em *transformers* e modelos S4 tradicionais, alcançando 90,27% de acurácia no conjunto *Breakfast* e 89,23% no *COIN*, demonstrando elevada capacidade de modelar dependências temporais complexas com menor custo computacional.

## 4 Um Estudo Exploratório Inicial

Neste capítulo, delineamos os passos envolvidos na condução da nossa pesquisa exploratória na convergência multimodal entre áudio e imagem. Cada decisão metodológica é fundamentada em uma cuidadosa revisão da literatura e em considerações teóricas, visando fornecer uma estrutura sólida para nossas investigações.

### 4.1 Visão Geral

A dualidade nos conjuntos de dados escolhidos reflete nossa intenção de explorar tanto relações básicas quanto complexas entre áudio e imagem. O comprometimento com a diversidade é central para nossa abordagem, visto que almejamos compreender as nuances nas interações multimodais. A arquitetura dos Variational Autoencoders (VAEs), exposta na Figura 4.1, foi então selecionada conscientemente, considerando não apenas a eficácia na aprendizagem de representações latentes (CAMUTO et al., 2021), mas também a flexibilidade necessária para lidar com a complexidade inerente aos dados multimodais.

A transformação  $G(x)$  do *encoder* é constituído de camadas convolucionais expostas na Equação 2.3. Já a transformação  $F(z)$  é composta por convoluções transpostas expostas na Equação 2.7. O processo de treinamento e avaliação é concebido como uma jornada iterativa. A qualidade das reconstruções e a diversidade nas amostras geradas servem como bússolas para ajustes contínuos. Essas métricas são cruciais para avaliar a eficácia dos modelos em reproduzir fielmente as características originais e, ao mesmo tempo, explorar a variabilidade nas relações intermodais.

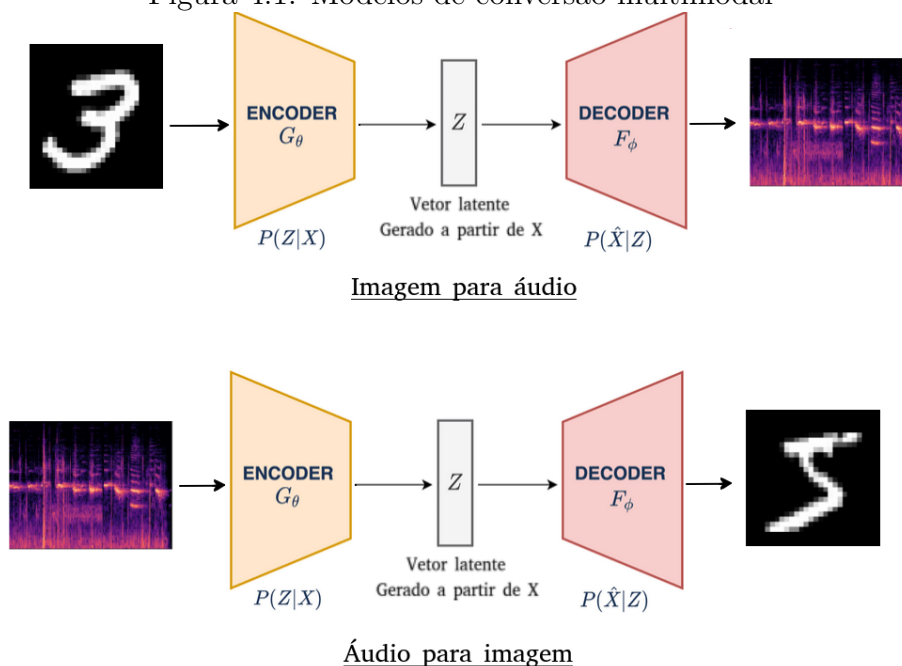
Nas seções subsequentes, cada componente será explorado com detalhes, oferecendo detalhes mais profundos sobre as escolhas estratégicas e os procedimentos específicos que nortearão nossa investigação na convergência multimodal.

### 4.2 Conjuntos de Dados

Em nossa abordagem, a seleção criteriosa de conjuntos de dados desempenha um papel crucial na capacidade de nossos modelos de aprender representações intermodais significativas. O primeiro conjunto de dados é uma fusão de áudios pronunciando os



Figura 4.1: Modelos de conversão multimodal



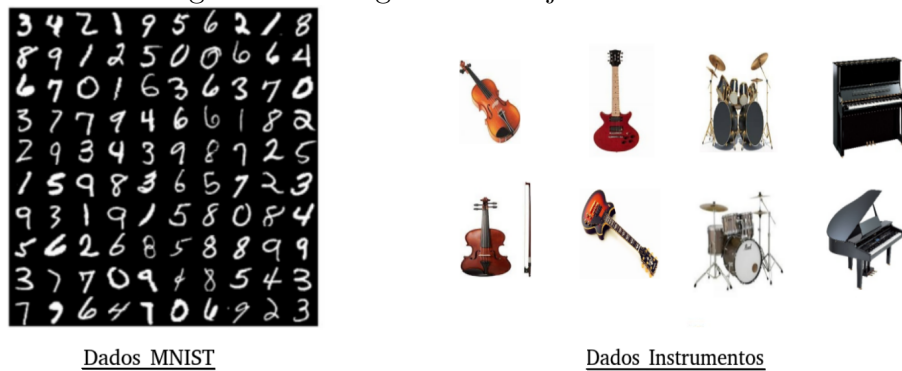
Fonte: Composição do autor

números de 0 a 9, capturados por diferentes locutores (JACKSON, 2016), juntamente com as imagens do MNIST (DENG, 2012). Cada exemplo neste conjunto é um par correspondente de áudio e imagem, onde áudios representando um mesmo dígito são pareados com a mesma representação visual. Essa diversidade intraclasse permite que o modelo aprenda relações sutis entre áudio e imagem associadas a cada número, contribuindo para uma convergência multimodal mais robusta.

O segundo conjunto de dados estende essa abordagem ao incorporar imagens e sons relacionados a instrumentos musicais. Semelhante ao primeiro conjunto, cada exemplo neste conjunto é composto por um par de áudio e imagem, onde diferentes instrumentos musicais são emparelhados com suas representações visuais correspondentes (MOHANTY, 2022; GPIOSENKA, 2022). Essa variedade de dados multimodais cria um ambiente desafiador e rico, permitindo que os modelos explorem nuances nas relações entre áudio e imagem em contextos musicais. Essa diversificação dos conjuntos de dados é essencial para nossa pesquisa, pois abrange uma gama ampla de relações intermodais que enriquecem a convergência entre áudio e imagem. Ambos os conjuntos podem ser observados na Figura 4.2

A divisão dos conjuntos de dados em conjuntos de treinamento, teste e validação foi realizada com uma proporção de 80/10/10, garantindo uma distribuição equilibrada e representativa para treinar e avaliar nossos modelos. Essa estratégia de divisão proporciona um ambiente de treinamento robusto e uma avaliação confiável do desempenho dos modelos,

Figura 4.2: Imagens dos conjuntos de dados



Fonte: Composição do autor

garantindo a generalização eficaz para dados não vistos.

No que concerne ao conjunto de dados do MNIST, contabilizamos um total de 3000 exemplos, com 300 exemplos para cada uma das classes (dígitos de 0 a 9). Os exemplos foram distribuídos entre 2400 para treinamento e 300 para os conjuntos de teste e validação. Já no segundo conjunto de dados, a totalidade é de 2705 exemplos, contendo 720 exemplos para as classes de Guitarra, Bateria e Violino, e 545 exemplos para a classe de Piano. Do total, 2163 são destinados ao treinamento e 271 reservados para cada um dos conjuntos de teste e validação.

### 4.3 Pré-processamento das Imagens

Aqui, adotamos uma abordagem de padronização para simplificar a tarefa de conversão multimodal. Isso inclui o redimensionamento de todas as imagens para um tamanho uniforme e a conversão para uma representação em tons de cinza. Essas etapas visam minimizar a complexidade do problema, tornando as imagens mais homogêneas e reduzindo o espaço de características a ser considerado pelos modelos. Essa estratégia facilita a aprendizagem das relações entre as modalidades de áudio e imagem, proporcionando um ambiente mais controlado para nossos modelos generativos.

Buscando a melhor representação possível das informações, utilizamos imagens com tamanho fixo de 128x128 pixels para ambos os conjuntos de dados. Isso irá garantir que o nosso modelo tenha informações suficientes para codificar de forma eficiente as características das imagens com um nível maior de complexidade estrutural.

## 4.4 Processamento de Áudio e Espectrogramas

O processamento de áudio no âmbito de redes neurais é extremamente similar ao de imagens. Para isso, uma etapa crucial é a transformação dos dados de áudio em uma representação visual. Os espectrogramas são representações visuais de sinais de áudio que destacam as frequências e a evolução temporal do som. Esse processo é obtido por meio da aplicação da STFT (DURAK; ARIKAN, 2003) aos sinais de áudio, cuja formulação matemática e definição já foram discutidos na Seção 2.9. Os espectrogramas resultantes são matrizes bidimensionais que representam o espectro de frequência ao longo do tempo.

Embora o espectrograma convencional já viabilize a extração de características auditivas por meio de modelos de aprendizado profundo, optamos pela utilização de uma variação específica conhecida como espectrograma de Mel (DÖRFLER; BAMMER; GRILL, 2017). Distinto do espectrograma convencional, o espectrograma de Mel se destaca por sua capacidade de reproduzir de maneira mais fiel a percepção auditiva humana das diferentes frequências. Para isso, primeiramente é feita uma conversão de escala das frequências do espectrograma obtidos com a STFT aplicando a seguinte transformação:

$$M(f) = 2595 \cdot \log_{10} \left( 1 + \frac{f}{700} \right). \quad (4.1)$$

Enquanto o espectrograma convencional representa linearmente as frequências, o espectrograma de Mel adota uma escala não linear, mais alinhada com a sensibilidade do ouvido humano. Isso se traduz em uma resolução mais precisa nas frequências mais baixas, onde a discriminação auditiva é mais aguçada.

No contexto específico do processamento de áudio para aprendizado de máquina, as vantagens do espectrograma de Mel se evidenciam. Ao mapear sinais de áudio em representações visuais mais alinhadas com a percepção humana, superamos barreiras na discriminação de características em diferentes frequências (KHURANA et al., 2023). Esta característica torna o espectrograma de Mel especialmente eficaz para extrair padrões discriminativos em tarefas de machine learning relacionadas a áudio. Essa riqueza de informações visuais proporciona uma base sólida para a extração de padrões sutis e complexos, contribuindo significativamente para o sucesso de aplicações intermodais, como a convergência entre áudio e imagem.

Os bancos de filtros Mel e os *bins* na STFT são componentes fundamentais na transformação do sinal de áudio em um espectrograma. Os bancos de filtros Mel, modelados a partir da sensibilidade auditiva humana, atuam como filtros passa-faixa, cada um capturando informações em uma faixa específica do espectro de frequência. Por sua vez, os *bins* na STFT representam intervalos discretos no domínio da frequência. A interação

entre esses elementos permite a extração de características cruciais do sinal de áudio. A manipulação do número de bancos de filtros Mel e *bins* na STFT não apenas influencia a representação espectral, mas também determina o tamanho final do espectrograma resultante, exercendo um controle direto sobre as dimensões da imagem final obtida.

Ao definir os parâmetros para a transformação do espectrograma de Mel, conseguimos regular o tamanho final da imagem, assegurando que esta coincida com a dimensão desejada. Para ambos os conjuntos de dados que apresentam imagens de 128x128 pixels, empregamos 128 bancos de filtros Mel, com 1024 *bins* para a STFT e saltos de janela de 337 unidades. Considerando uma taxa de amostragem padrão de 480000 Hz, esses parâmetros resultarão em um espectrograma de 128x128.

Ao incorporar o espectrograma de Mel em nossa metodologia, almejamos não apenas capitalizar suas diferenças fundamentais em relação ao espectrograma convencional, mas também explorar as vantagens que essas distinções conferem à análise de dados sonoros no contexto específico de aprendizado de máquina. Essa escolha estratégica enriquece nossa abordagem, proporcionando uma representação mais informativa e discriminativa, fundamental para a compreensão das intrincadas relações multimodais em nosso domínio de pesquisa.

## 4.5 Arquitetura do Modelo

A escolha do VAE como modelo fundamental em nossa abordagem decorre de uma série de vantagens notáveis em relação aos autoencoders tradicionais. Enquanto autoencoders convencionais são eficazes na compressão de informações e extração de características, VAEs introduzem uma camada adicional de complexidade por meio da modelagem probabilística. Essa característica permite uma representação mais robusta e contínua do espaço latente, oferecendo não apenas codificações eficientes, mas também uma noção mais clara da incerteza associada a cada ponto nesse espaço (CAMUTO et al., 2021).

A natureza probabilística do VAE também contribui significativamente para a geração de dados mais realistas. A capacidade de amostragem no espaço latente permite a criação de novas instâncias que não apenas se assemelham às presentes nos dados de treinamento, mas também capturam a variabilidade inerente às diferentes modalidades, enriquecendo assim a capacidade do modelo de representar dados multimodais.

A aplicação do VAE em nossa tarefa específica promete um desempenho aprimorado pela sua capacidade de modelar distribuições latentes de maneira mais eficaz. Na Figura 4.3, temos a arquitetura usada para treinamento em ambos os conjuntos de dados. Cada bloco de convolução exposto nas arquiteturas é composto por um bloco convolucional (vide Equação 2.3) ou convolucional transposta (vide Equação 2.7), ambas as alternativas

seguidas de uma camada de normalização de batch, também conhecida como Batch Normalization (BN). BN é uma técnica essencial em arquiteturas de redes neurais, para acelerar o treinamento e melhorar a estabilidade. Ela é aplicada a cada *mini-batch* de dados durante o treinamento e normaliza as ativações para ter média zero e variância unitária. A fórmula para a normalização de batch é dada por:

$$\text{BN}(x) = \gamma \left( \frac{x - \mu}{\sqrt{\sigma^2 + \epsilon}} \right) + \beta, \quad (4.2)$$

onde  $x$  é a entrada,  $\mu$  é a média do *mini-batch*,  $\sigma^2$  é a variância do *mini-batch*,  $\epsilon$  é um termo de suavização para evitar a divisão por zero, e  $\gamma$  e  $\beta$  são parâmetros de escala e viés aprendidos (IOFFE; SZEGEDY, 2015). Essa técnica contribui para a normalização dos gradientes durante o treinamento, melhorando a convergência da rede e permitindo o uso de taxas de aprendizado mais altas. Além disso, ela introduz parâmetros  $\gamma$  e  $\beta$ , que permitem à rede aprender a melhor escala e deslocamento para as ativações normalizadas. Ao fim de cada bloco, aplicamos a função de ativação Leaky ReLU (*Leaky Rectified Linear Unit (ReLU)*). Leaky ReLU é uma variação da função de ativação ReLU que introduz uma pequena inclinação para os valores negativos, em vez de ser estritamente zero. A sua equação é definida como:

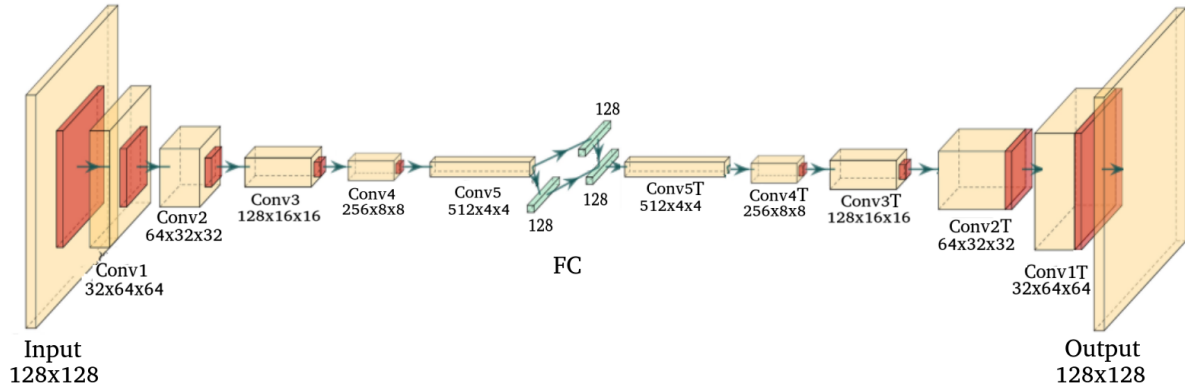
$$\text{LeakyReLU}(x) = \begin{cases} x, & \text{se } x \geq 0 \\ \alpha x, & \text{se } x < 0 \end{cases} \quad (4.3)$$

onde  $\alpha$  é uma pequena constante positiva que controla a inclinação para os valores negativos. A Leaky ReLU ajuda a mitigar o problema do “neurônio morto” que pode ocorrer em unidades ReLU tradicionais, permitindo um fluxo de gradiente mesmo para entradas negativas. Essa ativação é frequentemente usada em arquiteturas de redes neurais para evitar o problema de unidades inativas e promover uma convergência mais estável durante o treinamento.

A interpretação probabilística na codificação latente é particularmente crucial em nosso contexto de convergência multimodal, onde a complexidade das relações entre áudio e imagem exige uma representação latente que não apenas capture características distintivas, mas também considere a incerteza associada a essas relações.

Ao explorar o espaço latente probabilístico oferecido pelo VAE, nossa metodologia visa não apenas otimizar a representação de características relevantes, mas também fornecer uma base sólida para a análise das interações entre áudio e imagem. A capacidade do VAE de modelar de maneira mais eficiente as complexidades das distribuições de dados permitirá uma convergência mais significativa entre modalidades, potencializando assim nossa busca por uma compreensão aprofundada e rica das relações intermodais em nosso

Figura 4.3: Arquitetura da rede usada no treinamento



Fonte: Composição do autor

domínio de pesquisa.

## 4.6 Treinamento e Avaliação

No processo de treinamento, a escolha criteriosa da função de perda desempenha um papel central na capacidade do modelo de aprender representações significativas do conjunto de imagens geradas  $X$  para a saída  $Y$  de referência. Optamos por uma abordagem que combina o Mean Squared Error (MSE):

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (y_i - x_i)^2, \quad (4.4)$$

uma métrica convencional para avaliação de reconstruções, com a Maximum Mean Discrepancy (MMD):

$$\text{MMD}(F, p, q) = \|\mu_p - \mu_q\|_F^2, \quad (4.5)$$

onde  $p$  e  $q$  são distribuições de dados, e  $F$  é uma função de kernel. A MMD mede a diferença entre as distribuições  $p$  e  $q$  no espaço do kernel (TOLSTIKHIN et al., 2017). O MSE, ao penalizar a diferença entre cada reconstrução  $y_i$  e as instâncias referências  $x_i$ , assegura a fidelidade na reprodução das características fundamentais para cada um dos  $n$  exemplos do conjunto de treino. A inclusão do MMD, por sua vez, fortalece a capacidade do modelo em considerar a variabilidade intrínseca nas distribuições de dados, promovendo uma representação latente  $z$  mais rica e generalizável. A função de perda é dado pela soma dessas duas métricas das métricas apresentadas:

$$loss = MSE + MMD. \quad (4.6)$$

A avaliação efetiva do desempenho do modelo requer métricas criteriosamente selecionadas que capturem diversas facetas das tarefas propostas. Adotamos a MSE como métrica fundamental. Além da MSE, incorporamos o Structural Similarity Index Measure (SSIM):

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}. \quad (4.7)$$

como métrica adicional (NILSSON; AKENINE-MÖLLER, 2020). O SSIM avalia a similaridade estrutural entre as imagens geradas e as originais, considerando não apenas a diferença pixel a pixel, mas também aspectos de textura, luminância e contraste dada duas constantes  $C_1$  e  $C_2$ . Essa abordagem holística proporciona uma visão mais abrangente da fidelidade das reconstruções, considerando nuances perceptuais cruciais.

A escolha dessas métricas reflete a necessidade de avaliações abrangentes que vão além de simples medidas de erro, considerando a complexidade das relações multimodais. O MSE e o SSIM, combinados, oferecem uma perspectiva detalhada sobre a fidelidade e qualidade das reconstruções, elementos cruciais na validação e refinamento contínuo do modelo proposto.

## 4.7 Metodologia dos Experimentos

A condução de experimentos para validação de modelos multimodais requer uma exploração sistemática de diferentes configurações de parâmetros. Nesta fase exploratória do projeto, o objetivo foi avaliar o comportamento de *autoencoders variacionais* treinados separadamente nas modalidades auditiva e visual, sob condições controladas de treinamento. Com base nessa análise, buscamos entender se as representações latentes aprendidas eram estáveis, estruturadas e consistentes o suficiente para permitir reconstruções cruzadas entre domínios sensoriais.

Os principais hiperparâmetros avaliados incluíram o número de épocas de treinamento, a taxa de aprendizado (*learning rate*), o tamanho do *batch*, a estratégia de inicialização dos pesos e as técnicas de regularização. A variação controlada desses parâmetros permitiu avaliar seu impacto na convergência do modelo, na estabilidade durante o treinamento e na capacidade de generalização.

O número de épocas foi variado entre 10, 20, 30, 40, 50, 100 e 200, permitindo observar desde comportamentos de convergência precoce até os efeitos de *overfitting* em treinamentos mais longos. A taxa de aprendizado foi avaliada em valores entre  $5 \times 10^{-4}$  e  $5 \times 10^{-3}$ , com

incrementos de  $5 \times 10^{-4}$ , contemplando desde ajustes finos até passos de otimização mais agressivos. O tamanho do *batch* foi mantido fixo em 32, devido à limitação de memória da GPU utilizada.

Foram testadas diferentes estratégias de inicialização de pesos: Uniforme, normal, Xavier (GLOROT; BENGIO, 2010) e He (HE et al., 2015). Avaliamos sua influência na velocidade de convergência e na estabilidade de redes profundas. Também investigamos técnicas de regularização como L1, L2 (também conhecida como *weight decay*) e *dropout*, com taxas variando entre 0,1, 0,3 e 0,5. O objetivo dessas técnicas foi mitigar o *overfitting* e promover uma melhor generalização, especialmente em cenários com menor volume de dados ou alta variabilidade nas entradas.

Todos os experimentos foram conduzidos em uma estação de trabalho equipada com uma GPU RTX 3060 de 12GB garantindo reprodutibilidade e consistência entre os testes. O ambiente foi mantido estável e controlado, com os dados de entrada e os modelos executados localmente.

A Tabela 4.1 resume os principais parâmetros avaliados ao longo dos experimentos:

| Tabela 4.1: Resumo das Configurações Experimentais Avaliadas |                                                                          |
|--------------------------------------------------------------|--------------------------------------------------------------------------|
| Parâmetro                                                    | Valores Avaliados                                                        |
| Épocas de Treinamento                                        | 10, 20, 30, 40, 50, 100, 200                                             |
| Taxa de Aprendizado                                          | $5 \times 10^{-4}$ até $5 \times 10^{-3}$ (passo de $5 \times 10^{-4}$ ) |
| Tamanho do Batch                                             | 32                                                                       |
| Inicialização dos Pesos                                      | Uniforme, Normal, Xavier, He                                             |
| Regularização                                                | L1, L2, Dropout (0,1, 0,3, 0,5)                                          |

Após a execução dos experimentos, identificamos a configuração com melhor desempenho para os modelos de reconstrução unimodal. Os melhores resultados foram obtidos com 40 épocas, taxa de aprendizado de  $5 \times 10^{-5}$ , *batch size* de 32, inicialização dos pesos via Xavier e regularização com *dropout* de 0,1. Essa configuração promoveu uma convergência estável e evitou *overfitting*, ao mesmo tempo em que manteve boa fidelidade nas reconstruções. A Tabela 4.2 apresenta o resumo da configuração final.

Tabela 4.2: Melhor Configuração para os Modelos de Reconstrução

| Parâmetro               | Valor              |
|-------------------------|--------------------|
| Épocas de Treinamento   | 40                 |
| Taxa de Aprendizado     | $5 \times 10^{-5}$ |
| Tamanho do Batch        | 32                 |
| Inicialização dos Pesos | Xavier             |
| Regularização           | Dropout (0,1)      |

Esta abordagem metodológica permitiu uma análise abrangente da sensibilidade dos modelos aos principais hiperparâmetros e forneceu uma base sólida para o desenho da



arquitetura X-Net apresentada nos capítulos seguintes.

## 4.8 Resultado e Análise do Treinamento

Na abordagem da conversão de áudio em imagem, observamos progressos substanciais na capacidade do modelo em representar, de maneira fiel, as características intrínsecas do áudio por meio de imagens reconstruídas. Os resultados quantitativos, observados na Tabela 4.3, avaliados por métricas como *Mean Squared Error* e *Structural Similarity Index*, destacam uma precisão alta para o MNIST na reprodução das características fundamentais.

Tabela 4.3: Métricas para conversão de áudio para imagem

| Conjunto de Dados | MSE    | SSIM |
|-------------------|--------|------|
| MNIST             | 0,0187 | 0,95 |
| Instrumentos      | 0,1119 | 0,82 |

Fonte: Composição do autor

A visualização dos resultados é de extrema importância para a compreensão da qualidade das reconstruções obtidas. Exemplificamos essa melhoria por meio de representações visuais que ilustram a convergência entre os dados de áudio originais e as imagens geradas pelo modelo na Figura 4.4. Esses exemplos visuais proporcionam uma percepção intuitiva e tangível das melhorias conquistadas na tarefa de conversão áudio em imagem, evidenciando a eficácia do modelo na preservação das nuances acústicas.

Figura 4.4: Imagens geradas a partir de um conjunto de áudios



MNIST

Instrumentos

Fonte: Composição do autor

Esses avanços na conversão áudio em imagem sinalizam um progresso em direção à convergência multimodal desejada, evidenciando o potencial do modelo em proporcionar reconstruções visualmente representativas a partir de dados de áudio.

No contexto da conversão de imagem em áudio, os resultados indicam um progresso promissor na fidelidade da representação sonora a partir das imagens reconstruídas. As métricas quantitativas, MSE e SSIM observadas na Tabela 4.4, mostram uma performance

Tabela 4.4: Métricas para conversão de imagem para áudio

| Conjunto de Dados | MSE    | SSIM  |
|-------------------|--------|-------|
| MNIST             | 0,0412 | 0,924 |
| Instrumentos      | 0,2106 | 0,771 |

Fonte: Composição do autor

otimizada, destacando a eficácia do método, especialmente quando comparada ao fluxo inverso de áudio para imagem

Os resultados dos modelos de conversão foram bastante positivos, evidenciando a eficácia dos métodos empregados em ambas as direções. As métricas quantitativas como MSE e SSIM demonstraram um desempenho robusto na fidelidade das representações tanto sonoras quanto visuais. Com esses resultados encorajadores, realizamos uma análise das relações intrínsecas entre esses modelos, buscando compreender melhor as interdependências e sinergias que podem ser exploradas para aprimorar ainda mais o conhecimento sobre o processamento multimodal.

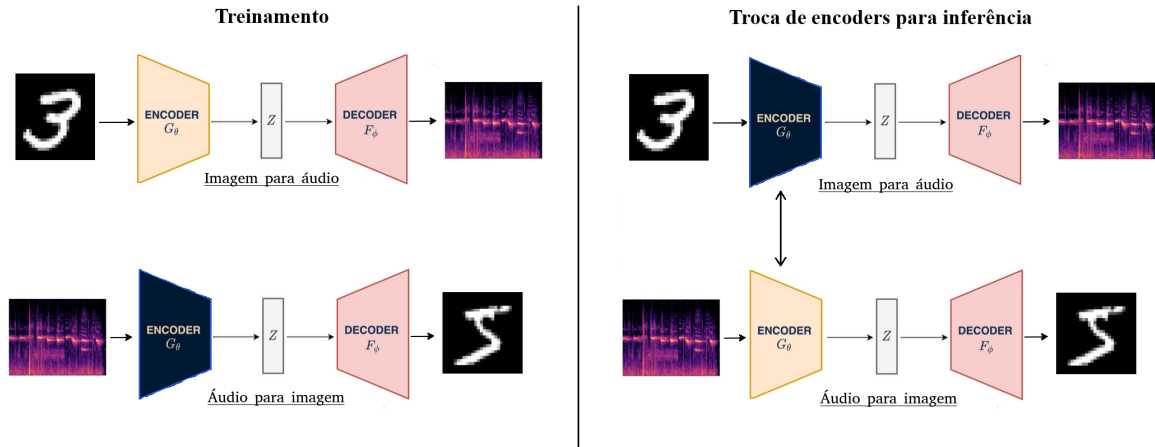
Durante os testes conduzidos, observamos um resultado notável. Ao combinar o *encoder* do modelo de imagem para áudio com o *decoder* do modelo de áudio para imagem, conforme ilustrado na Figura 4.5, conseguimos gerar imagens de saída que apresentam uma notável semelhança com imagens reais (Figura 2.6). Esta observação indica que as características aprendidas pelo *encoder* do modelo de imagem são de certa forma análogas às características aprendidas pelo *encoder* de áudio. Em outras palavras, o *decoder* do modelo áudio para imagem consegue produzir imagens estruturalmente similares as reais mesmo quando o *encoder* utilizado pertence ao modelo oposto, sugerindo um aprendizado cruzado e uma relação significativa entre os atributos extraídos de ambas as modalidades. Exemplos auditivos podem ser conferidos no material complementar disponível no repositório do projeto<sup>1</sup>.

A importância desta descoberta reside na demonstração de que as redes neurais, ao aprenderem a representar informações de áudio e imagem de maneira eficaz, acabam por desenvolver representações internas que capturam características essenciais e compartilháveis entre as duas modalidades. Este fenômeno não apenas valida a eficácia dos modelos treinados, mas também sugere uma capacidade intrínseca das redes neurais em entender e relacionar informações complexas e multifacetadas provenientes de diferentes domínios sensoriais.

Baseados nessa descoberta, propomos uma nova arquitetura bioinspirada no processamento multimodal no nosso cérebro. Esta nova abordagem capitaliza essa capacidade de

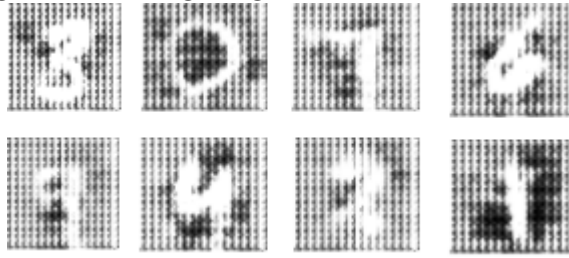
<sup>1</sup>Repositório GitHub: <https://github.com/GuFurlanetto/X-Net-Material-Complementar>

Figura 4.5: Troca de encoders para inferência



Fonte: Composição do autor

Figura 4.6: Imagens geradas com um VAE misto



Fonte: Composição do autor

aprendizado cruzado entre modalidades, visando melhorar não apenas a qualidade das conversões entre áudio e imagem, mas também explorar novas aplicações em reconhecimento de padrões multimodais e síntese de informações heterogêneas. Ao inspirar-se no cérebro humano, que integra eficientemente informações de múltiplos sentidos para formar percepções coesas, esperamos que nossa arquitetura avance significativamente na fronteira do processamento de sinais multimodais, abrindo novas possibilidades em áreas como reconhecimento de objetos em vídeos, síntese de fala a partir de imagens, entre outras aplicações.

# 5 X-Net: Arquitetura de Convergência Audiovisual Bioinspirada

Neste capítulo, detalhamos a proposta de uma arquitetura inovadora que visa integrar modalidades sensoriais de maneira inspirada no processamento cerebral humano. A arquitetura é projetada para imitar os mecanismos de integração sensorial observados no cérebro, com o objetivo de melhorar a análise e classificação de dados multimodais.

## 5.1 Visão Geral

A arquitetura proposta para processamento multimodal, observada na Figura 5.1, é inspirada nos mecanismos de integração sensorial do cérebro humano, particularmente no modo como o cérebro processa informações auditivas e visuais. Este modelo é projetado para explorar a interação entre diferentes modalidades sensoriais, com o objetivo de melhorar a compreensão e a classificação de dados complexos, como vídeos.

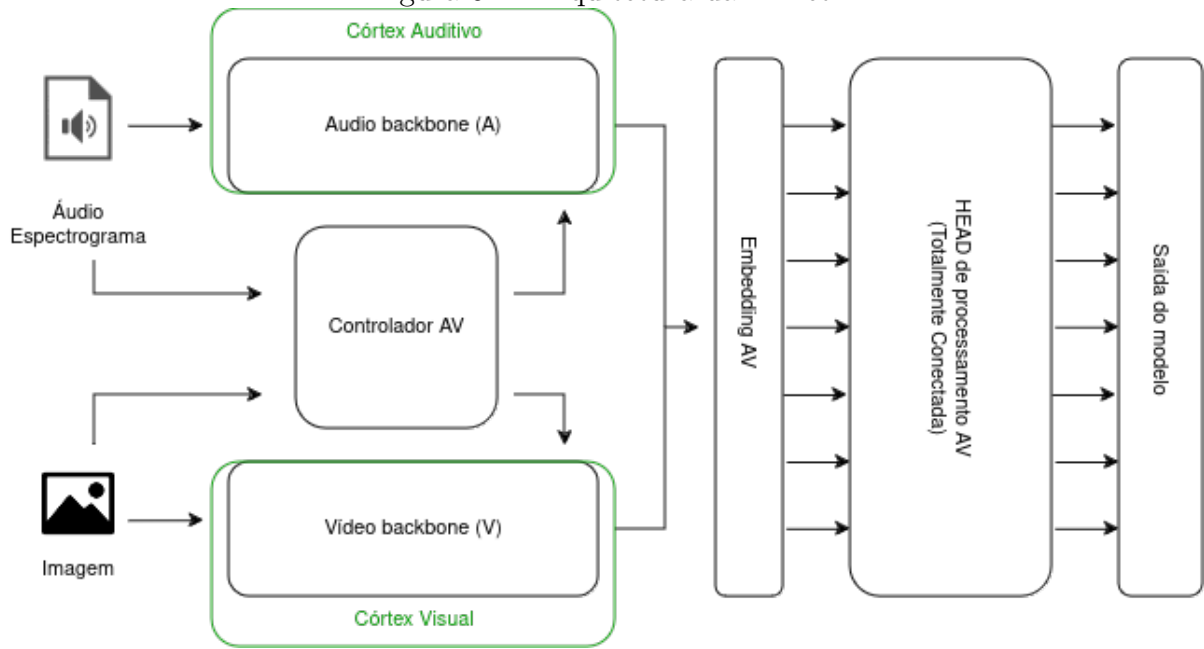
A arquitetura é composta por três componentes principais:

1. *Backbones* pré-treinadas no processamento unimodal, análogos aos córtex visual e auditivo.
2. Um controlador AV (áudio-visual) para integração multimodal entre as modalidades.
3. Etapa final de processamento específico para a tarefa objetivo (*HEAD*).

O uso das *backbones* pré-treinados ViT (DOSOVITSKIY et al., 2020) para áudio e VideoMAE (TONG et al., 2022) para vídeo permite a extração de características avançadas e robustas de cada modalidade.

A inspiração para esta arquitetura vem das descobertas sobre o processamento multimodal no cérebro. Estudos (SCHEEF et al., 2009; BESLE et al., 2008) indicam a união das modalidades nos estágios iniciais de processamento. Essa integração precoce permite que o cérebro tenha uma percepção mais coesa e contextualizada dos eventos

Figura 5.1: Arquitetura da X-Net



Fonte: Composição do Autor

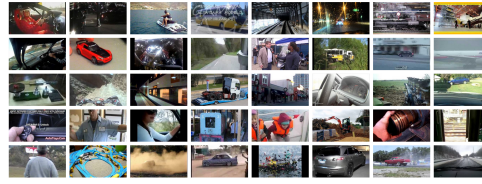
ao seu redor. Por exemplo, o efeito McGurk, onde a percepção do som é alterada pela visualização simultânea de movimentos labiais incongruentes, ilustra como a percepção auditiva e visual estão intrinsecamente conectadas.

## 5.2 Conjunto de Dados

Para a validação da nossa arquitetura bioinspirada, utilizaremos três conjuntos de dados: YouTube-8M (ABU-EL-HAIJA et al., 2016), UCF101 (SOOMRO; ZAMIR; SHAH, 2012) e COIN (TANG DAJUN DING, 2019). Estes conjuntos de dados foram escolhidos por suas características distintas e pela amplitude de desafios que apresentam, permitindo uma avaliação robusta e abrangente da eficácia do modelo proposto. A seguir, iremos explorar cada conjunto de dados individualmente:

- **YouTube-8M:** Este é um dos maiores conjuntos de dados de vídeos anotados, contendo aproximadamente 8 milhões de vídeos, o que corresponde a mais de 500 mil horas de vídeo (como pode ser observado na Figura 5.2), com representações pré-computadas de características visuais e auditivas. Cada vídeo, dividido em segmentos de 5 segundos, é anotado com uma ou mais das 3862 etiquetas, cobrindo uma ampla gama de categorias. Este conjunto foi utilizado como fonte para coleta de exemplos audiovisuais sem rótulo, visando o treinamento autossupervisionado de alguns dos módulos intermediários da arquitetura.

Figura 5.2: YouTube - 8M



Fonte: Adaptado de Abu-El-Haija et al. (2016)

- **UCF101:** O conjunto UCF101, ilustrado na Figura 5.3, é um *benchmark* consolidado para tarefas de reconhecimento de ações em vídeo. Ele contém 13.320 vídeos distribuídos em 101 categorias de ações humanas, como “arremesso de bola”, “tocando violão” e “corrida”. Os vídeos são curtos, geralmente com duração inferior a 10 segundos, e apresentam um alto grau de variação de cenário, iluminação e ponto de vista. Este *dataset* fornece fluxos audiovisuais sincronizados, permitindo uma análise conjunta das modalidades. Sua estrutura segmentada e a diversidade de atividades o tornam particularmente adequado para avaliação de modelos multimodais em tarefas de classificação.

Figura 5.3: Conjunto UCF101



Fonte: Adaptado de Soomro, Zamir e Shah (2012)

- **COIN:** Este conjunto de dados, observado na Figura 5.4, contém mais de 11 mil vídeos tutoriais de instrução em uma variedade de domínios, como culinária, esportes e manutenção doméstica, totalizando cerca de 476 horas de conteúdo. Cada vídeo é segmentado e anotado com 180 diferentes tipos de tarefas, abrangendo 12 domínios principais. Sua estrutura hierárquica de tarefas o torna ideal para análise de etapas e ações compostas, desafiando o modelo a compreender contextos multimodais complexos.

Cada um desses conjuntos de dados fornece desafios únicos que ajudam a avaliar diferentes aspectos da arquitetura proposta. O YouTube-8M, com sua vasta quantidade

Figura 5.4: Conjunto COIN



Fonte: Adaptado de Tang Dajun Ding (2019)

de dados e diversidade de categorias, foi utilizado para tarefas de pré-treinamento autossupervisionado, testando a capacidade de generalização dos módulos internos. O UCF101 oferece uma base padronizada para comparação com abordagens unimodais e multimodais em tarefas de ação. Já o COIN, com seus vídeos tutoriais segmentados, fornece um cenário mais complexo, no qual a integração das modalidades auditiva e visual é essencial para a correta inferência das tarefas.

A combinação desses conjuntos de dados serve como base para validar a arquitetura bioinspirada em múltiplos cenários, proporcionando uma variedade de desafios encontrados em aplicações reais de visão computacional e processamento multimodal.

### 5.3 Arquitetura

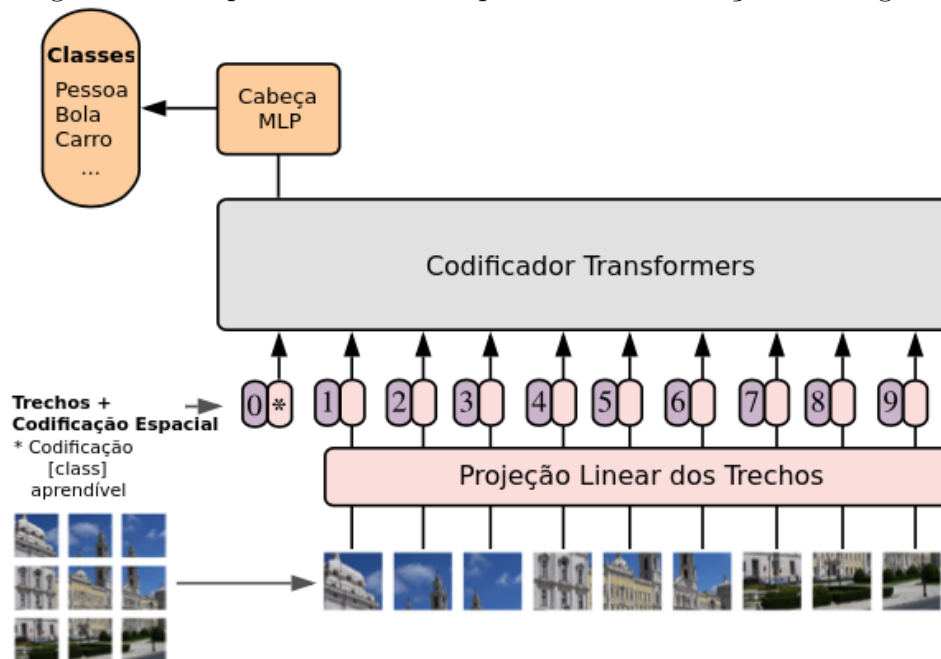
A arquitetura, observada na Figura 5.1, para processamento multimodal bioinspirada combina dados auditivos e visuais de maneira sinérgica, inspirando-se em mecanismos do sistema nervoso central humano para melhorar o processamento multimodal.

A arquitetura começa com duas *backbones* especializadas, uma para cada modalidade sensorial. Estas *backbones* são análogas aos córtex auditivo e visual no cérebro humano:

- **Backbone de Áudio (A):** Para a modalidade auditiva, utilizamos uma arquitetura baseada em *Vision Transformer* (ViT) (DOSOVITSKIY et al., 2020), adaptada para o processamento de espectrogramas de Mel. Esta escolha permite que o modelo capture dependências de longo alcance nas representações espectrais, fornecendo uma codificação rica e contextualizada dos sinais auditivos. O ViT é treinado diretamente sobre imagens espectrográficas com dimensões  $224 \times 224$ , permitindo a extração de *embeddings* de alto nível compatíveis com os mecanismos de fusão multimodal propostos. A Figura 5.5 ilustra o funcionamento interno do ViT, destacando as etapas de projeção dos *patches* de imagem, soma com codificações posicionais e

posterior processamento por camadas *transformer*, que produzem um vetor de classe final para cada entrada.

Figura 5.5: Arquitetura do ViT aplicada à classificação de imagens

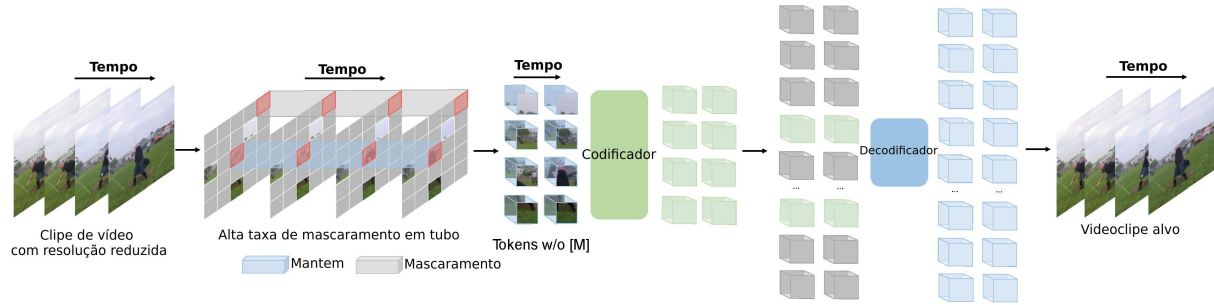


Fonte: Adaptado de Dosovitskiy et al. (2020)

- **Backbone de Vídeo (V):** Para a modalidade visual, utilizamos o modelo VideoMAE (TONG et al., 2022). Este modelo emprega *autoencoders* com mascaramento espaço-temporal, sendo treinado em grandes conjuntos de vídeos. Sua capacidade de processar sequências de *frames* e capturar a dinâmica temporal dos movimentos o torna particularmente eficaz em tarefas que envolvem reconhecimento de ações, além de fornecer representações latentes de alto nível, fundamentais para a integração multimodal proposta. A Figura 5.6 apresenta o fluxo de pré-treinamento do VideoMAE, no qual cliques de vídeo são convertidos em cubos de *tokens*, parcialmente mascarados ao longo das dimensões espaciais e temporais. Os tokens restantes são processados por um codificador, e o decodificador tenta reconstruir o clipe original, incentivando o aprendizado de padrões dinâmicos ricos e informativos. Durante o treinamento, o VideoMAE é otimizado por meio de uma tarefa de reconstrução, na qual grandes porções do vídeo são mascaradas e o modelo é forçado a prever os conteúdos ausentes. Essa abordagem autossupervisionada promove o aprendizado de representações latentes ricas e generalizáveis. No entanto, durante a inferência, o decodificador é descartado, e apenas o codificador é utilizado para extrair as *embeddings* finais dos vídeos. Isso reduz significativamente a complexidade computacional, mantendo a expressividade das representações aprendidas.



Figura 5.6: Processo de pré-treinamento do modelo VideoMAE com mascaramento espaço-temporal



Fonte: Adaptado de Tong et al. (2022)

Essas *backbones* realizam o processamento inicial das entradas de áudio e vídeo, extraindo características específicas de cada modalidade de forma independente.

## 5.4 Controlador AV

A inovação central da nossa arquitetura é o Controlador AV, que imita as áreas de integração sensorial presentes no cérebro humano, como o Sulco Temporal Superior (STS) e o Giro Temporal Superior (STG). Conforme discutido em maior detalhe na Seção 3.9, essas regiões desempenham papel fundamental na fusão entre estímulos auditivos e visuais. Inspirado nesse mecanismo biológico, o Controlador AV foi projetado para promover a integração precoce entre as modalidades, enriquecendo o processamento de cada fluxo unimodal e permitindo uma interpretação conjunta mais robusta.

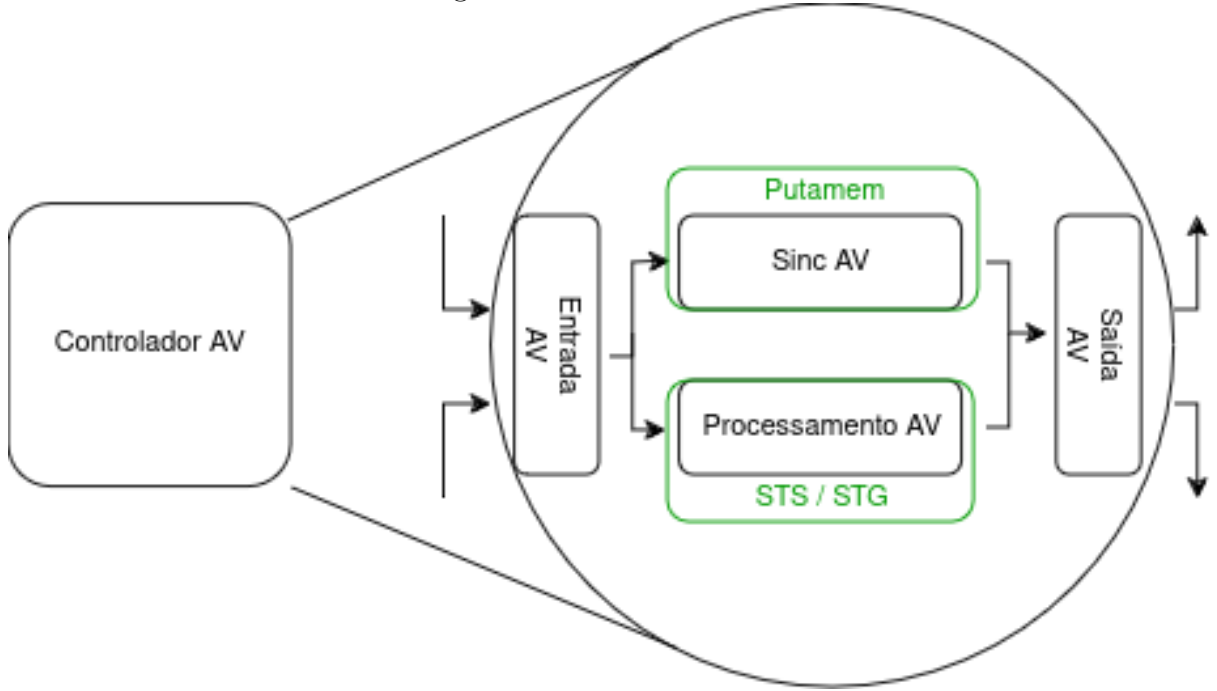
Sua estrutura modular, ilustrada na Figura 5.7, é composta por quatro blocos principais: *Entrada AV*, *SincAV*, *Processamento AV* e *Saída AV*. Cada um desses blocos desempenha um papel específico no fluxo de dados, desde o pré-processamento das modalidades até a geração de uma representação multimodal unificada, que será usada para modular a saída produzida pelas *backbones*.

### Entrada AV

A entrada AV é responsável por organizar e preparar os dados provenientes das modalidades auditiva e visual para os módulos internos do Controlador AV. Este bloco opera em paralelo ao processamento inicial realizado pelas *backbones*, assegurando que os dados de entrada estejam devidamente alinhados e normalizados para permitir a aprendizagem de representações multimodais compartilhadas.

Para o áudio, o sinal bruto é convertido em espectrogramas de Mel bidimensionais.

Figura 5.7: Controlador AV



Fonte: Composição do Autor

Especificamente, são utilizados 128 bancos de filtros Mel e uma STFT com 1024 *bins*, com saltos de 512 unidades. A taxa de amostragem padrão de 16kHz é mantida em todos os casos, garantindo compatibilidade com os vídeos de entrada. O resultado dessa transformação é um espectrograma log-Mel com resolução final redimensionada para 224×224 pixels, compatível com a arquitetura ViT utilizada como *backbone* auditiva.

Para o vídeo, os quadros são extraídos a uma taxa fixa de 1 quadro por segundo e redimensionados para 224×224 pixels. Cada amostra é composta por uma sequência de 16 quadros consecutivos, organizados em blocos temporais. Essa estrutura permite que os modelos sejam capazes de capturar padrões dinâmicos ao longo do tempo, como movimentos, transições e sincronia entre ações e sons. No caso do VideoMAE, essas sequências de quadros são empilhadas e organizadas em tensores tridimensionais com dimensão  $T \times H \times W \times C$  (onde  $T = 16$ ,  $H = W = 224$ , e  $C = 3$ ), seguindo o formato de entrada esperado pelo modelo. Internamente, o VideoMAE divide cada sequência em pequenos *patches* espaciais e temporais, que são embarcados em tokens e processados por meio de camadas transformadoras com atenção espaço-temporal mascarada, permitindo a extração de representações ricas e informativas ao longo do tempo.

O bloco *SincAV*, descrito mais adiante, requer acesso explícito à sequência temporal dos quadros para detectar variações de sincronicidade entre os fluxos. Já o *Processador AV* opera com instâncias estáticas, visando o aprendizado cruzado entre as representações extraídas de ambas as modalidades.

O processo de normalização das entradas também é padronizado: as imagens (quadros e espectrogramas) são convertidas para a faixa  $[0, 1]$  e posteriormente normalizadas para o intervalo  $[-1, 1]$ , utilizando média e desvio padrão de 0,5. Essa normalização é compatível com os requisitos das arquiteturas baseadas em ViT e garante que os dados estejam no mesmo domínio de distribuição esperado durante o treinamento dos módulos internos.

Dessa forma, o bloco de entrada AV garante que todas as instâncias multimodais sejam processadas de maneira uniforme e coordenada, assegurando a integridade e a coerência dos dados ao longo das etapas subsequentes do Controlador AV.

## SincAV

O modelo *SincAV*, observado na Figura 5.8, foi desenvolvido com o objetivo de identificar a sincronia entre áudio e vídeo em segmentos multimodais. Inspirado na atuação do Putâmen e do STS no cérebro humano, regiões especializadas na integração temporal de estímulos sensoriais, o módulo busca aprender representações capazes de capturar alinhamentos e desalinhamentos entre os sinais auditivos e visuais.

Sua arquitetura é modular, composta por dois codificadores independentes: um dedicado à entrada de vídeo e outro ao espectrograma de áudio. Os codificadores processam os dados separadamente, extraindo vetores latentes que são posteriormente concatenados e utilizados em tarefas de predição de sincronização. Essa separação permite que cada caminho aprenda características específicas da sua modalidade, ao passo que a concatenação promove uma fusão multimodal eficiente.

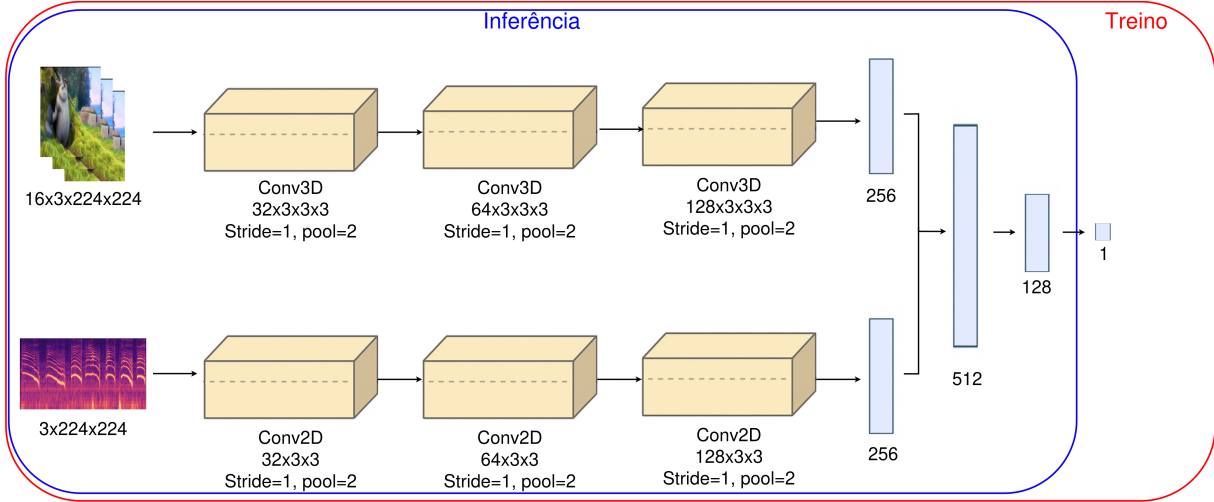
O conjunto de dados utilizado para treinamento foi construído a partir de vídeos reais com duração de cinco segundos. Esses exemplos foram divididos em duas categorias principais:

- **Casos positivos**, onde o áudio e o vídeo mantêm sua sincronização natural;
- **Casos negativos**, onde o áudio é deslocado temporalmente para frente ou para trás, ou substituído por trilhas sonoras sem relação semântica com as imagens correspondentes.

A arquitetura dos codificadores baseia-se em redes convolucionais, sendo convoluções 3D para o fluxo de vídeo e convoluções 2D para o espectrograma. Cada codificador é composto por blocos convolucionais seguidos de camadas de *Batch Normalization*, funções de ativação ReLU e operações de *MaxPooling*. Ao final da extração de características, os vetores latentes concatenados.

A etapa final consiste em uma cabeça preditiva dedicada à tarefa de classificação binária, cujo objetivo é identificar se os sinais de áudio e vídeo estão sincronizados. Essa

Figura 5.8: Arquitetura SincAV



Fonte: Composição do Autor

cabeça é composta por camadas lineares e uma função de ativação final do tipo sigmoide, que produz a probabilidade da classe positiva (sincronizado). Durante o treinamento, a função de perda utilizada é a Binary Cross Entropy (BCE), adequada para tarefas de decisão binária (GOODFELLOW; BENGIO; COURVILLE, 2016):

$$\mathcal{L}_{\text{BCE}}(y, \hat{y}) = -(y \cdot \log(\hat{y}) + (1 - y) \cdot \log(1 - \hat{y})), \quad (5.1)$$

onde  $y \in \{0, 1\}$  representa o rótulo verdadeiro (sincronizado ou não) e  $\hat{y} \in [0, 1]$  representa a probabilidade predita pelo modelo.

Ao final do treinamento, espera-se que o módulo seja capaz de identificar deslocamentos sutis de sincronia entre os fluxos multimodais, funcionando como uma instância artificial do sistema biológico de sincronização sensorial presente no cérebro humano.

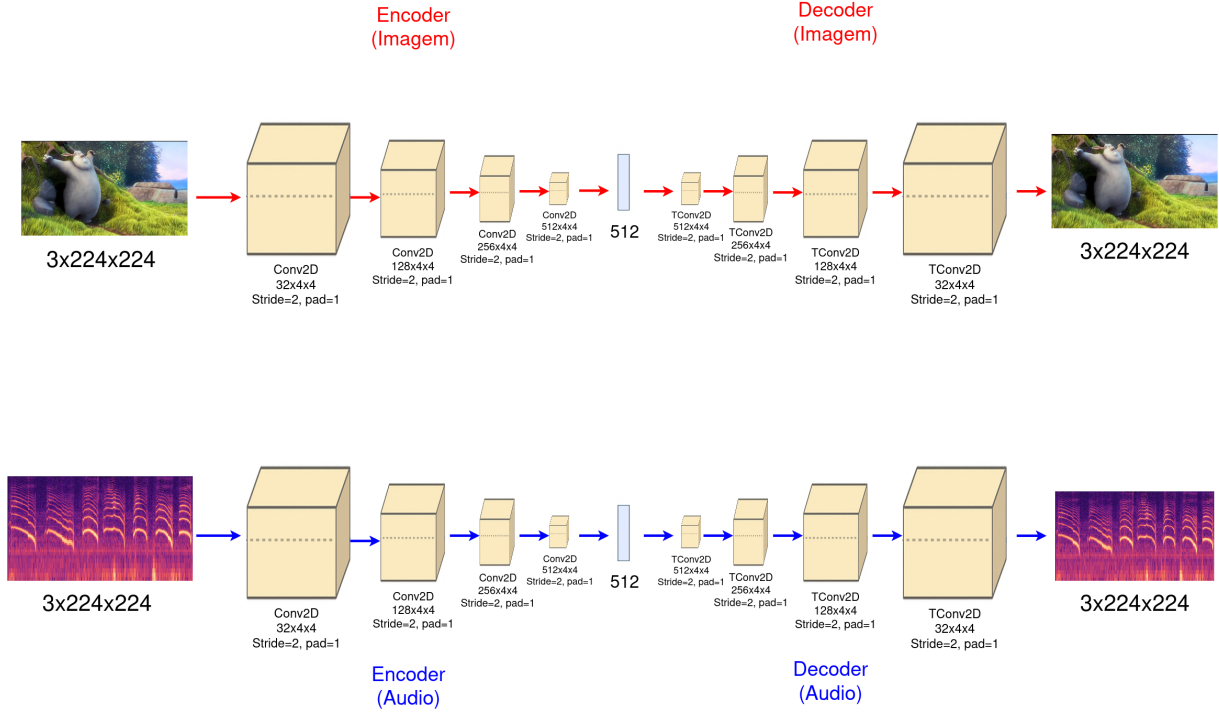
## Processamento AV

A etapa de Processamento AV é responsável pela modelagem cruzada entre as modalidades, permitindo a conversão de representações auditivas em representações visuais e vice-versa. Inspirado em mecanismos biológicos observados no Sulco Temporal Superior e no Giro Temporal Superior, este módulo atua como um elo entre os fluxos sensoriais, promovendo a integração e transposição das informações capturadas pelas modalidades.

Sua estrutura é baseada em dois *Variational Autoencoders* (VAEs) especializados: um para o fluxo de áudio e outro para o fluxo de vídeo. Cada VAE é composto por um codificador (*encoder*) e um decodificador (*decoder*), os quais são treinados para reconstruir

a entrada. Essa configuração, ilustrada na Figura 5.9, permite que o sistema aprenda um espaço latente, no qual representações auditivas e visuais possam ser mapeadas de forma consistente.

Figura 5.9: Arquitetura Processamento AV (Treinamento)



Fonte: Composição do Autor

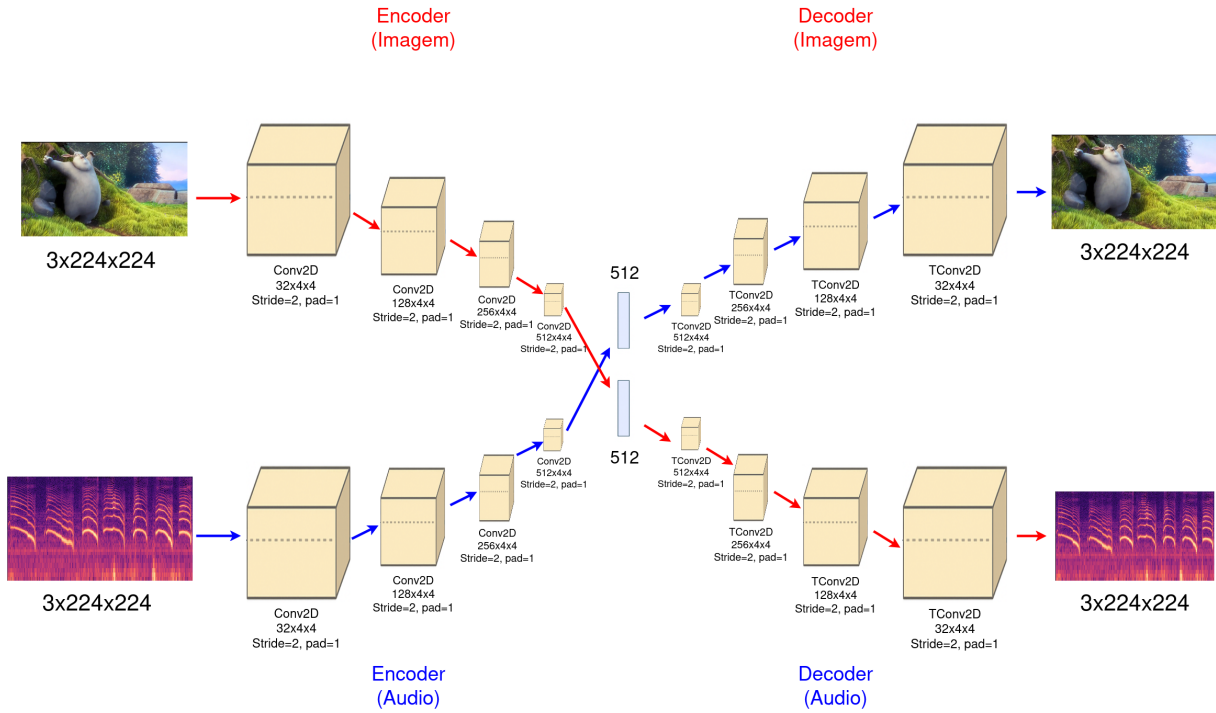
Durante o treinamento, optamos por um processo incremental que difere da abordagem adotada nos modelos de conversão explorados na etapa anterior do projeto. Enquanto nos modelos de conversão buscava-se diretamente a transposição entre modalidades, aqui priorizamos uma estratégia mais prática e eficiente, focada inicialmente no treinamento unimodal dos dois *autoencoders variacionais*. Ou seja, cada VAE foi otimizado para reconstruir sua própria modalidade (áudio ou vídeo) a partir de seus dados originais. Essa decisão foi motivada por questões práticas: a coleta, o processamento e o alinhamento dos dados são substancialmente mais simples no regime unimodal, além de proporcionar maior estabilidade e velocidade durante os estágios iniciais de aprendizado. A ideia central, no entanto, permanece a mesma: explorar a capacidade dos *encoders* de capturar representações latentes informativas, estruturadas e potencialmente compartilháveis entre os domínios. Para isso, utilizamos uma combinação de funções de perda: o *Mean Squared Error* (MSE) entre a entrada reconstruída e a real, o *Structural Similarity Index* (SSIM) para capturar aspectos perceptuais da fidelidade estrutural da reconstrução, e a *Maximum Mean Discrepancy* (MMD) para regularizar o espaço latente, forçando a distribuição aprendida a se aproximar de uma distribuição normal multivariada:

$$\mathcal{L}_{\text{total}} = \text{MSE} + (1 - \text{SSIM}) + \lambda \cdot \text{MMD}, \quad (5.2)$$

onde o termo  $\lambda$  controla a normalização do espaço latente à distribuição gaussiana. Essa combinação de métricas incentiva o modelo a preservar a estrutura das amostras e manter uma representação latente coerente.

Na fase de inferência, conforme mostrado na Figura 5.10, o Processador AV realiza uma troca entre os caminhos de decodificação: o encoder de vídeo é conectado ao decoder de áudio e vice-versa. Essa estratégia de reconstrução cruzada foi motivada pelos resultados obtidos na fase exploratória, na qual observou-se que os codificadores aprendiam representações latentes compartilháveis entre as modalidades, mesmo quando utilizados de forma intercambiável.

Figura 5.10: Arquitetura Processamento AV (Inferência)



Fonte: Composição do Autor

Durante a inferência, o fluxo auditivo é convertido em uma representação visual e incorporado no caminho de processamento da imagem, e o mesmo ocorre para o fluxo visual. Essa troca de informações entre os caminhos sensoriais visa simular a capacidade do cérebro de inferir propriedades visuais a partir de sons, ou características sonoras a partir de estímulos visuais, fenômeno amplamente documentado em estudos de integração sensorial (MCGURK; MACDONALD, 1976; BEAUCHAMP et al., 2004).

Esse cruzamento entre as modalidades é o elemento central da proposta do Contro-

lador AV e da própria arquitetura *X-Net*. Ao permitir que os fluxos de entrada sejam reinterpretados e reconstruídos pela modalidade oposta, o modelo busca emular uma das características mais fundamentais do processamento multimodal biológico: a capacidade de representação cruzada entre domínios sensoriais complementares.

## Saída AV

A *Saída AV* é o bloco responsável por consolidar as informações extraídas pelos módulos *SincAV* e *Processador AV* e transformá-las em representações latentes que serão utilizadas para modulação das *backbones* unimodais. Em vez de realizar uma simples concatenação dos *embeddings*, o processamento da saída é conduzido de forma estruturada, por meio da combinação espacial dos *embeddings* com as saídas geradas em cada modalidade.

Mais especificamente, o vetor latente produzido pelo módulo *SincAV* é expandido espacialmente por meio de uma operação de *broadcast*, resultando em um mapa de ativação com as mesmas dimensões espaciais das imagens geradas pelo *Processador AV*. Esse mapa de *embeddings* é então concatenado com os dados reconstruídos (o espectrograma de áudio gerado e o frame de vídeo correspondente) formando um par de tensores multicanal multimodais.

Cada um desses tensores fundidos é então processado por uma pequena rede convolucional especializada (uma para cada modalidade) cuja função é integrar espacialmente os sinais de sincronização com o conteúdo reconstruído. Essas redes possuem estrutura simples e eficiente: consistem em duas camadas convolucionais com *kernels*  $3 \times 3$  e *padding* 1, seguidas de funções de ativação ReLU. A primeira camada utiliza 64 filtros e a segunda 128, permitindo uma progressiva extração de padrões multimodais locais. Após essa fusão convolucional, aplica-se uma operação de *global average pooling*, que reduz os mapas de ativação resultantes a vetores latentes compactos. Esses vetores finais de modulação capturam de forma condensada o conteúdo multimodal relevante extraído por todo o Controlador AV.

Esses vetores capturam de forma condensada o conteúdo multimodal relevante extraído por todo o Controlador AV. Dessa forma, a Saída AV opera como elo entre os módulos internos do Controlador e os caminhos principais de classificação, viabilizando uma fusão multimodal profunda, coerente e extensível.

### 5.4.1 Modulação das Backbones por FiLM

A etapa final do processamento realizado pelo Controlador AV consiste na geração de duas *embeddings* latentes de modulação: uma voltada para o fluxo visual e outra para o fluxo auditivo. Essas representações são obtidas por meio da fusão entre os *embeddings* de

sincronização e os dados reconstruídos pelo *Processador AV*, conforme descrito nas seções anteriores. O objetivo dessas *embeddings* é condicionar dinamicamente o comportamento das *backbones* unimodais, promovendo uma integração contextual entre os sinais auditivo e visual.

Para efetivar essa integração, adotamos o mecanismo de *Feature-wise Linear Modulation* (FiLM), uma técnica que permite ajustar representações neurais com base em sinais condicionais externos. O funcionamento do FiLM é formalmente descrito por (PEREZ et al., 2018):

$$\text{FiLM}(x) = \gamma \cdot x + \beta, \quad (5.3)$$

onde  $x \in \mathbb{R}^{B \times D}$  representa uma matriz de ativações, com  $B$  indicando o tamanho do *batch* e  $D$  a dimensionalidade da *embedding* extraída de uma das *backbones*. Os vetores  $\gamma, \beta \in \mathbb{R}^{B \times D}$  são parâmetros aprendidos condicionalmente a partir das *embeddings* de modulação. Essa operação permite que o contexto multimodal fornecido pelo Controlador AV influencie diretamente as representações finais produzidas por cada modalidade.

Em nossa implementação, a modulação é aplicada após o processamento completo de cada *backbone*, diretamente sobre as *embeddings* finais extraídas de vídeo e áudio:

- A saída da **backbone visual** (VideoMAE), representada como  $\mathbf{e}_v \in \mathbb{R}^{D_b}$ , é modulada utilizando os parâmetros FiLM gerados a partir da *embedding*  $\mathbf{z}_v \in \mathbb{R}^{D_m}$ , produzida pelo Controlador AV.
- De maneira análoga, a saída da **backbone auditiva** (ViT sobre espectrogramas),  $\mathbf{e}_a \in \mathbb{R}^{D_b}$ , é modulada utilizando os parâmetros gerados a partir de  $\mathbf{z}_a \in \mathbb{R}^{D_m}$ .

Aqui,  $D_b$  representa a dimensionalidade das *embeddings* finais extraídas pelas *backbones*, enquanto  $D_m$  corresponde à dimensionalidade das *embeddings* de modulação geradas pelo Controlador AV.

As funções de modulação são implementadas por redes lineares específicas que transformam cada *embedding* de modulação  $\mathbf{z}$  em vetores  $\gamma$  e  $\beta$  com a mesma dimensionalidade de  $\mathbf{e}$ , garantindo compatibilidade no processo de ajuste. A aplicação da FiLM ocorre diretamente sobre as *embeddings* extraídas, resultando em versões moduladas  $\tilde{\mathbf{e}}_v$  e  $\tilde{\mathbf{e}}_a$ , enriquecidas com o contexto multimodal.

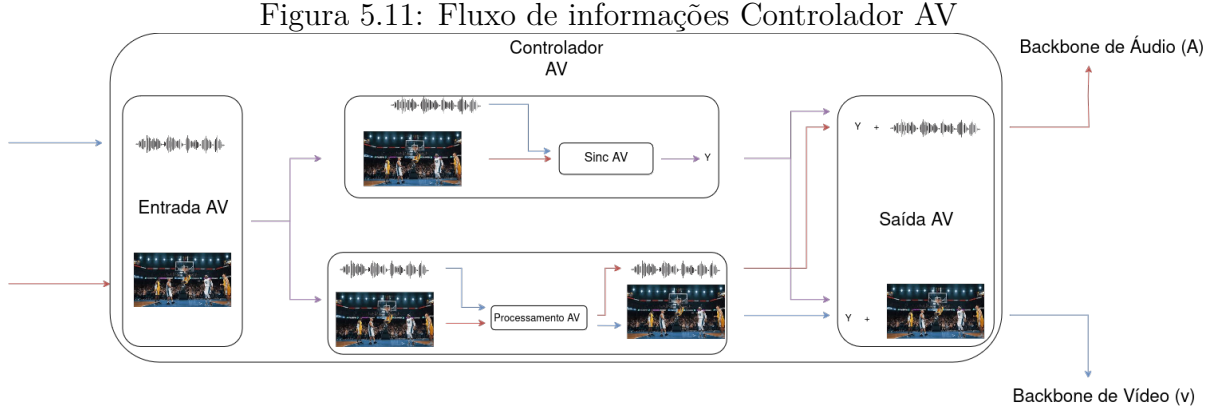
Essa abordagem é inspirada em mecanismos biológicos de modulação cortical, nos quais regiões sensoriais distintas são influenciadas por sinais provenientes de áreas integradoras por meio de vias de modulação seletiva, sem perda da especialização local.

Por fim, essas representações moduladas são concatenadas para formar a *embedding* multimodal consolidada:



$$\mathbf{e}_{\text{final}} = [\tilde{\mathbf{e}}_v \parallel \tilde{\mathbf{e}}_a] \in \mathbb{R}^{2D_b}, \quad (5.4)$$

a qual é então encaminhada à *head* de classificação, responsável por produzir a predição final do modelo. O fluxo de informações do controlador pode ser observado na Figura 5.11.



Fonte: Composição do Autor

## 5.4.2 Treinamento da arquitetura

O treinamento da arquitetura X-Net foi estruturado em três etapas principais: o pré-treinamento dos módulos internos do *Controlador AV* usando aprendizado autossupervisionado, seguido pelo ajuste fino supervisionado da arquitetura integrada para a tarefa de classificação multimodal de vídeos.

### Treinamento do módulo SincAV

O módulo *SincAV* foi treinado com o objetivo de detectar sincronização ou dessincronização entre os fluxos auditivo e visual. Para isso, adotamos uma estratégia de aprendizado autossupervisionado, sem a necessidade de anotações manuais.

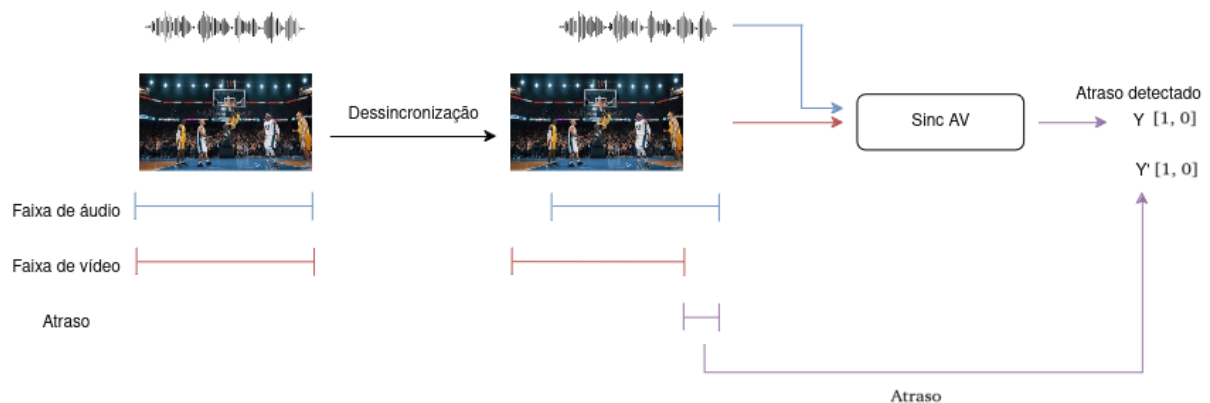
A base de treinamento foi construída a partir de vídeos reais, assumindo que o áudio original desses vídeos encontra-se naturalmente sincronizado com a imagem. Esses exemplos formaram os **casos positivos**. A partir deles, geramos automaticamente os **casos negativos** por meio de duas estratégias: (i) introdução de deslocamentos temporais artificiais no áudio, variando entre -4 e +4 segundos; e (ii) substituição da trilha sonora por segmentos de áudio aleatórios sem relação semântica com o conteúdo visual. Esse processo está ilustrado na Figura 5.12.

A escolha do intervalo de deslocamento entre -4 e +4 segundos foi motivada por dois

critérios principais. Primeiramente, esse valor é próximo da duração típica dos cliques utilizados, o que garante que o trecho de áudio deslocado ainda possua alguma parte relacionada ao fluxo visual. Isso evita ambiguidade quanto à fonte do áudio, ao mesmo tempo em que ainda desafia o modelo a detectar dessincronizações perceptíveis. Em segundo lugar, deslocamentos menores poderiam não ser suficientemente perceptíveis para caracterizar uma dessincronização clara, enquanto deslocamentos muito maiores poderiam se tornar indistinguíveis de uma substituição completa da faixa de áudio. Como o problema foi modelado como uma tarefa de classificação binária (sincronizado vs. dessincronizado), deslocamentos mais amplos não oferecem vantagem adicional e poderiam até prejudicar a clareza dos exemplos negativos.

O modelo foi treinado para classificar cada par como sincronizado ou dessincronizado, a partir da entrada de um espectrograma de Mel e uma sequência de quadros de vídeo. Essa tarefa binária permitiu ao SincAV aprender representações latentes sensíveis ao alinhamento temporal entre as modalidades.

Figura 5.12: Treinamento SincAV



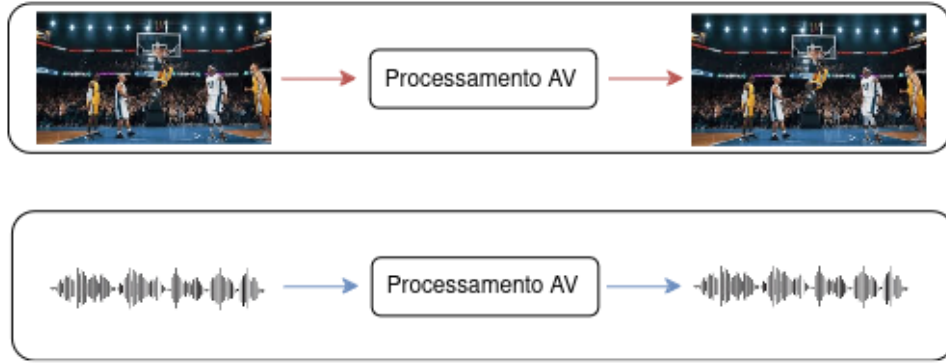
Fonte: Composição do Autor

## Treinamento do módulo Processador AV

O *Processador AV* foi treinado com base em duas instâncias unimodais separadas: um VAE para áudio e outro para vídeo. Como observado na Figura 5.13, cada autoencoder foi treinado para reconstruir a sua própria modalidade a partir de dados de entrada reais. Essa escolha permite maior estabilidade no treinamento, facilitando a convergência de cada fluxo e permitindo que os encoders aprendam representações latentes estruturadas antes da integração cruzada na fase de inferência.

A função de perda combinou o erro quadrático médio (MSE), a discrepância de média máxima (MMD) e o índice estrutural de similaridade (SSIM), conforme descrito na Equação 5.3.

Figura 5.13: Treinamento Processador AV



Fonte: Composição do Autor

### Treinamento supervisionado da X-Net

Após o pré-treinamento dos módulos internos, realizamos o treinamento supervisionado da arquitetura completa para a tarefa de classificação de vídeos. Nesta fase, o módulo *Controlador AV* é utilizado em modo inferência: o *SincAV* e o *Processador AV* são mantidos congelados, e suas saídas são usadas para gerar os vetores de modulação que influenciam as *backbones*.

Realizamos *fine-tuning* completo dos seguintes componentes durante essa fase:

- As duas *backbones* (VideoMAE e ViT);
- As camadas de modulação FiLM (responsáveis por aplicar os parâmetros  $\gamma$  e  $\beta$ );
- A *head* final de classificação.

Essa decisão visa adaptar os extratores de características às particularidades dos dados utilizados, ao mesmo tempo em que se preserva o conhecimento multimodal aprendido nos módulos internos do Controlador AV.

A representação final da amostra, composta pela concatenação dos *embeddings* finais das duas *backbones*, é passada à *head* de classificação, composta por camadas lineares e função de ativação *softmax*. A função de perda utilizada é a Cross-Entropy, definida da seguinte forma (GOODFELLOW; BENGIO; COURVILLE, 2016):

$$\mathcal{L}_{\text{CE}}(y, \hat{y}) = - \sum_{c=1}^C y_c \log(\hat{y}_c), \quad (5.5)$$

onde  $C$  é o número de classes,  $y_c$  é o valor binário do rótulo verdadeiro para a classe  $c$ , e  $\hat{y}_c$  é a probabilidade predita pelo modelo.

---

Essa configuração final de treinamento permite que a arquitetura X-Net aprimore suas representações de forma contextualizada, integrando a informação auditiva e visual de maneira coerente e sinérgica, conforme observado em sistemas biológicos de percepção multimodal.

## 6 Análise dos Resultados

Com o treinamento concluído para os três blocos principais da arquitetura (SincAV, Processador AV e X-Net) passamos agora à análise quantitativa e qualitativa dos resultados obtidos. Neste capítulo, apresentamos os experimentos realizados, as métricas avaliadas, as visualizações geradas e uma comparação final com trabalhos correlatos da literatura.

### 6.1 Ambiente de Execução

Todos os experimentos foram conduzidos em uma estação de trabalho equipada com os seguintes recursos computacionais:

- GPU: NVIDIA RTX 4070 Super com 12GB de VRAM;
- CPU: AMD Ryzen 7 5700X (8 cores, 16 threads);
- Memória RAM: 32GB DDR4;
- Sistema Operacional: Ubuntu 22.04 LTS;
- Frameworks: PyTorch 2.7.1, CUDA 12.9, Optuna para busca de hiperparâmetros.

A execução dos treinamentos foi realizada em modo *offline*, com todos os conjuntos de dados pré-processados e armazenados localmente para reduzir o tempo de Input/Output (I/O) durante os experimentos.

### 6.2 Resultados do Módulo SincAV

O módulo SincAV foi avaliado com base em sua capacidade de distinguir entre segmentos multimodais sincronizados e dessincronizados. Inicialmente, experimentamos formulá-lo como um problema de regressão, tentando estimar o deslocamento temporal exato entre as modalidades. No entanto, essa abordagem mostrou-se limitada, sobretudo pela dificuldade em garantir que todos os deslocamentos introduzidos fossem de fato perceptíveis ou relevantes do ponto de vista semântico. Como consequência, observamos instabilidade

durante o treinamento e baixa correlação entre os valores preditos e a percepção real de sincronização.

Considerando o objetivo final do módulo (produzir uma representação latente que capture indícios de alinhamento temporal entre as modalidades), optamos por reformular o problema como uma tarefa de classificação binária. Essa mudança simplificou a modelagem dos dados e possibilitou melhor controle na construção dos pares positivos (fluxos sincronizados) e negativos (fluxos com deslocamento de áudio ou trilhas não relacionadas).

Com essa reformulação, o modelo atingiu um bom desempenho, alcançando uma acurácia de 81% no conjunto de validação. A qualidade das predições do modelo pode ser visualizada qualitativamente nos vídeos incluídos no material complementar disponível no repositório do projeto <sup>1</sup>. Cada vídeo representa uma instância fora do conjunto de treinamento, contendo exatamente a mesma cena visual. Em um dos casos, o áudio original foi mantido; no outro, foi antecipado em alguns segundos. O nome do arquivo indica a predição do modelo, cuja saída é um valor contínuo entre 0 e 1: quanto mais próximo de 1, maior a confiança de que há sincronização entre os fluxos.

Essa visualização reforça a robustez do modelo na identificação de alinhamento temporal em cenários reais, mesmo diante de alterações sutis de sincronismo.

A escolha dos hiperparâmetros de treinamento foi realizada por meio de busca bayesiana utilizando a biblioteca *Optuna*. Esse procedimento automatizado permitiu explorar eficientemente o espaço de configurações possíveis, otimizando o desempenho no conjunto de validação. Os melhores resultados foram obtidos com o otimizador AdamW, taxa de aprendizado inicial de  $1 \times 10^{-3}$ , *weight decay* de  $1 \times 10^{-4}$  e *LR Scheduler* do tipo *ExponentialLR* com fator  $\gamma = 0,95$ . O treinamento foi conduzido por até 100 épocas, com critério de *early stopping* após 10 épocas sem melhora, e tamanho de *batch* de 16 amostras. A Tabela 6.1 apresenta os hiperparâmetros finais adotados.

Tabela 6.1: Hiperparâmetros utilizados no treinamento do SincAV

| Parâmetro                                   | Valor                             |
|---------------------------------------------|-----------------------------------|
| Otimizador                                  | AdamW                             |
| Taxa de aprendizado inicial                 | $1 \times 10^{-3}$                |
| Decaimento de pesos ( <i>weight decay</i> ) | $1 \times 10^{-4}$                |
| <i>LR Scheduler</i>                         | ExponentialLR ( $\gamma = 0,95$ ) |
| <i>batch size</i>                           | 16                                |
| Épocas máximas                              | 100                               |
| <i>Early stopping</i>                       | 10 épocas de paciência            |
| Baralhamento dos dados de treinamento       | Sim                               |

<sup>1</sup>Repositório GitHub: <https://github.com/GuFurlanetto/X-Net-Material-Complementar>

## 6.3 Resultados do Processador AV

O módulo Processador AV foi avaliado em tarefas de reconstrução unimodal, separadamente para os fluxos de entrada visual (imagem  $\rightarrow$  imagem) e auditivo (áudio  $\rightarrow$  áudio). Como métricas de avaliação, utilizamos *Structural Similarity Index Measure* (SSIM) e *Peak Signal-to-Noise Ratio* (PSNR), que avaliam a fidelidade perceptual e a qualidade da reconstrução em termos de ruído, respectivamente.

Os resultados apresentados na Tabela 6.2 demonstram que o Processador AV foi capaz de reconstruir ambas as modalidades de estrutura e baixo nível de ruído, mesmo considerando um conjunto de dados substancialmente maior do que o utilizado nos modelos de conversão anteriores. A similaridade dos resultados com aqueles obtidos nos experimentos iniciais reforça a robustez da arquitetura utilizada.

Tabela 6.2: Resultados do Processador AV nas reconstruções unimodais

| Métrica | Áudio $\rightarrow$ Áudio | Imagem $\rightarrow$ Imagem |
|---------|---------------------------|-----------------------------|
| SSIM    | 62,23%                    | 71,5%                       |
| PSNR    | 19,5 dB                   | 21,5 dB                     |

O processo de treinamento foi direto, uma vez que a arquitetura do Processador AV herda a base dos modelos de conversão explorados na primeira fase deste trabalho, exigindo apenas ajustes mínimos para adaptação ao novo volume de dados. A escolha final dos hiperparâmetros foi realizada por meio de validação periódica a cada 60 etapas, com parada antecipada após 10 iterações sem melhora na métrica de validação. O otimizador AdamW foi utilizado com *weight decay*, combinado a um agendador de taxa de aprendizado do tipo OneCycleLR. A Tabela 6.3 resume os parâmetros finais utilizados:

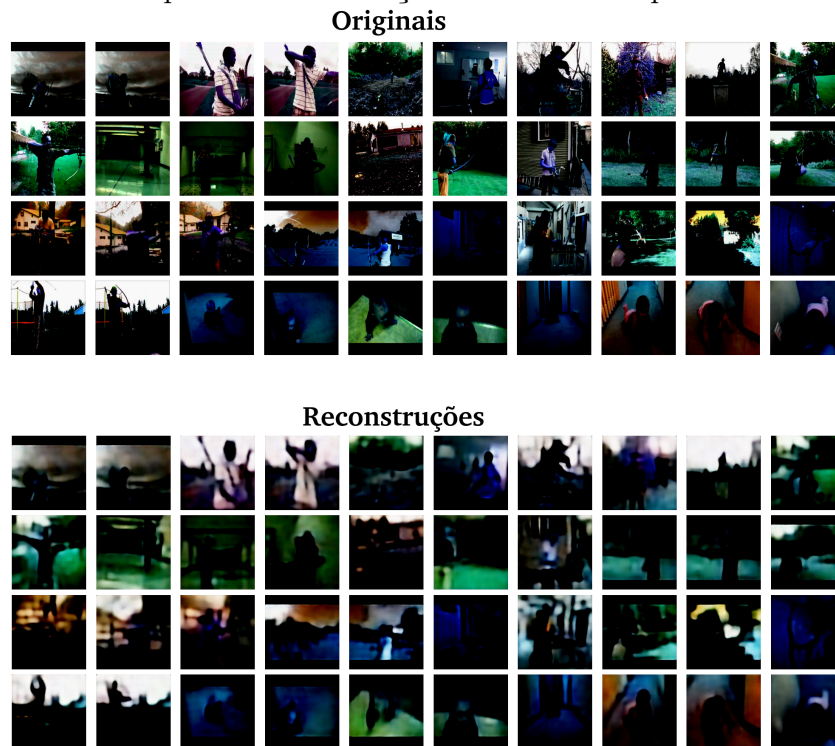
Tabela 6.3: Hiperparâmetros utilizados no treinamento do Processador AV

| Parâmetro                                   | Valor                                                                          |
|---------------------------------------------|--------------------------------------------------------------------------------|
| Otimizador                                  | AdamW                                                                          |
| Taxa de aprendizado inicial                 | $1 \times 10^{-3}$                                                             |
| Decaimento de pesos ( <i>weight decay</i> ) | $1 \times 10^{-4}$                                                             |
| <i>LR Scheduler</i>                         | OneCycleLR ( $\text{max\_lr} = 1 \times 10^{-3}$ , $\text{pct\_start} = 0,3$ ) |
| <i>batch size</i>                           | 64                                                                             |
| Épocas máximas                              | 100                                                                            |
| <i>Early stopping</i>                       | 10 épocas de paciência                                                         |
| Baralhamento dos dados de treinamento       | Sim                                                                            |

Como visualização qualitativa, a Figura 6.1 apresenta amostras de reconstruções visuais obtidas pelo modelo a partir de entradas reais. No caso do áudio, exemplos de espectrogramas reconstruídos são exibidos na Figura 6.2.

Além da avaliação direta das reconstruções, realizamos também uma análise qualitativa de

Figura 6.1: Exemplos de reconstrução visual obtidos pelo Processador AV



Fonte: Composição do Autor

interpolação no espaço latente. A ideia é verificar se o espaço aprendido entre *embeddings* é suave e semanticamente coerente, permitindo transições graduais entre diferentes entradas.

Para isso, selecionamos pares de amostras da modalidade visual (imagens de entrada) e realizamos uma interpolação linear entre seus *embeddings* latentes, com pesos variando de 0 a 1. As representações interpoladas foram então decodificadas de volta ao domínio da imagem. A Figura 6.3 apresenta uma sequência ilustrando essa transição entre duas entradas reais.

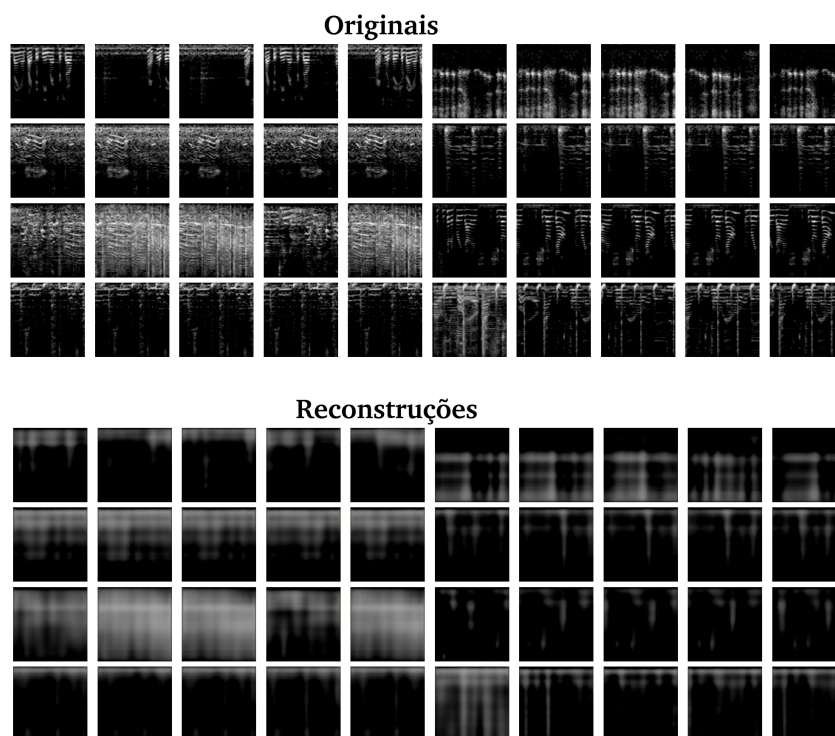
Observa-se que o modelo foi capaz de gerar imagens intermediárias coerentes, com transições suaves e plausíveis entre os atributos das duas imagens originais. Esse comportamento reforça a hipótese de que o espaço latente aprendido pela arquitetura é estruturado e semanticamente informativo, o que é um requisito importante para aplicações posteriores de manipulação ou transferência multimodal.

## 6.4 Resultados da X-Net

A Tabela 6.4 apresenta os resultados finais da arquitetura completa X-Net na tarefa de classificação de vídeos, comparando duas configurações: com e sem o uso do *Controlador*



Figura 6.2: Exemplos de reconstrução de espectrogramas obtidos pelo Processador AV



Fonte: Composição do Autor

Figura 6.3: Interpolação no espaço latente visual aprendida pelo Processador AV



Fonte: Composição do Autor

AV. Os experimentos foram realizados sobre os conjuntos de dados COIN e UCF101, e os resultados reportados referem-se à acurácia Top-1 obtida em cada cenário.

Os resultados indicam que o uso do *Controlador AV* teve impacto positivo e consistente no desempenho da arquitetura, com melhorias superiores a 5 pontos percentuais em ambos os conjuntos. Isso demonstra que a modulação condicional das *backbones*, guiada pelas *embeddings* temporais e reconstrutivas do *Controlador AV*, contribui significativamente para a extração de representações multimodais mais discriminativas.

Embora a X-Net não tenha superado os modelos de ponta em termos de acurácia absoluta, especialmente no conjunto COIN, seus resultados permanecem competitivos entre as abordagens que exploram fluxos audiovisuais (A+V). Enquanto métodos como o IMD-B utilizam múltiplas modalidades, incluindo texto (A+V+T), e arquiteturas como o MA-LMM e o VideoMamba operam exclusivamente sobre o fluxo visual (V), a X-Net

Tabela 6.4: Comparação de modelos para a tarefa de classificação de vídeo

| Modelo                                         | Modalidades | UCF101       | COIN         |
|------------------------------------------------|-------------|--------------|--------------|
| IMD-B-(GPT-text) (2024) (ALFASLY et al., 2024) | A+V+T       | <b>99,3%</b> | -            |
| AvSlowFast (2020) (XIAO et al., 2020)          | A+V         | 87,0%        | -            |
| AVTS (2018) (KORBAR; TRAN; TORRESANI, 2018)    | A+V         | 85,7%        | -            |
| Multisensory (2018) (OWENS; EFROS, 2018)       | A+V         | 82,1%        | -            |
| C3D (2018)* (HUANG et al., 2018)               | V           | 82,3%        | -            |
| R(2+1)D (2019)* (XU et al., 2019)              | V           | 72,4%        | -            |
| MA-LMM (2024) (HE et al., 2024)                | V           | -            | <b>93,2%</b> |
| VideoMamba (2024) (LI et al., 2024)            | V           | -            | 90,4%        |
| TranS4mer (2023) (ISLAM et al., 2023)          | V           | -            | 89,3%        |
| <b>X-Net (sem Controlador AV)</b>              | A+V         | 75,5%        | 52,3%        |
| <b>X-Net (com Controlador AV)</b>              | A+V         | 81,7%        | 59,1%        |

\* A resolução das imagens de entrada (*frames*) não é especificada pelos autores nos artigos correspondentes.

Fonte: Composição do Autor

propõe uma abordagem bioinspirada que explora de forma sinérgica a interação entre áudio e vídeo.

Comparando com o AvSlowFast, que combina múltiplos caminhos de processamento visual (lento e rápido) com um caminho dedicado ao áudio e integra as informações em múltiplos níveis hierárquicos, a X-Net adota uma pipeline mais compacta e modular. Apesar de o AvSlowFast alcançar 87,0% no UCF101, sua complexidade e dependência de integração em diferentes estágios contrastam com a nossa estratégia de fusão tardia via modulação FiLM, que prioriza interpretabilidade e reutilização de backbones pré-treinadas. Em relação ao AVTS, que utiliza sincronização áudio-vídeo como tarefa de pré-treinamento e depois transfere as representações para classificação, a X-Net incorpora o sinal de sincronização como componente ativo durante a inferência, modulando diretamente as *backbones*. Isso resulta em um desempenho competitivo no UCF101 (81,7% contra 85,7%), considerando a diferença na escala de pré-treinamento. Quando comparada ao Multisensory, que realiza fusão precoce com convoluções 3D diretamente nos sinais brutos, a X-Net se beneficia de extratores de características especializados (VideoMAE e ViT) antes da fusão. Essa escolha favorece a modularidade, mas pode sacrificar parte da sinergia capturada pela fusão precoce, explicando a diferença de 0,4 ponto percentual no UCF101 (82,1% contra 81,7%).

O ganho obtido com a introdução do *Controlador AV* valida a proposta de integração sensorial com modulação FiLM, evidenciando seu potencial para reforçar a representatividade de *embeddings* multimodais com maior interpretabilidade e menor dependência de

grandes volumes de dados supervisionados. Tais resultados indicam que a X-Net se apresenta como uma arquitetura promissora para aplicações multimodais, ao unir desempenho competitivo e estrutura modular, dois elementos frequentemente em tensão na literatura atual.

Apesar da melhora consistente proporcionada pelo *Controlador AV*, observa-se uma diferença significativa de desempenho entre os dois conjuntos de dados utilizados. No conjunto UCF101, a X-Net alcançou 81,7% de acurácia, enquanto no COIN o valor máximo obtido foi de 59,1%. Essa discrepância pode ser atribuída principalmente à maior complexidade estrutural do COIN, cujos vídeos são, em média, significativamente mais longos e compostos por múltiplas etapas semânticas. Ao contrário do UCF101, que contém vídeos curtos e bem segmentados de ações humanas isoladas, o COIN apresenta tarefas compostas, como receitas ou procedimentos, que exigem entendimento de contexto, ordem temporal e composição multimodal mais refinada.

Outro fator que contribuiu para a diferença de desempenho foi a estratégia adotada para a inferência, que seguiu o mesmo procedimento descrito no artigo do VideoMAE (TONG et al., 2022): a amostragem de um único segmento de vídeo por amostra, representado por um bloco de 16 quadros consecutivos. Embora eficiente computacionalmente, essa abordagem é limitada quando aplicada a vídeos longos, como os do COIN, pois pode capturar apenas uma fração não representativa da atividade principal. Dessa forma, mesmo que a modulação multimodal promova uma melhor fusão entre os fluxos de entrada, a ausência de uma estratégia de agregação temporal mais abrangente penaliza a capacidade da arquitetura de compreender eventos distribuídos ao longo do tempo.

O treinamento da X-Net utilizou otimização por meio do algoritmo AdamW, com *weight decay* e estratégia de *OneCycleLR* para o ajuste dinâmico da taxa de aprendizado. O processo foi realizado durante 20 épocas, com critério de *early stopping* configurado para 5 épocas. Por fim, a Tabela 6.5 resume os principais hiperparâmetros empregados no treinamento da arquitetura.

Tabela 6.5: Hiperparâmetros utilizados no treinamento da X-Net

| Parâmetro                                   | Valor                                                                          |
|---------------------------------------------|--------------------------------------------------------------------------------|
| Otimizador                                  | AdamW                                                                          |
| Taxa de aprendizado inicial                 | $5 \times 10^{-5}$                                                             |
| Decaimento de pesos ( <i>weight decay</i> ) | $1 \times 10^{-4}$                                                             |
| <i>LR Scheduler</i>                         | OneCycleLR ( $\text{max\_lr} = 5 \times 10^{-5}$ , $\text{pct\_start} = 0,1$ ) |
| <i>batch size</i>                           | 4                                                                              |
| Épocas máximas                              | 20                                                                             |
| <i>Early stopping</i>                       | 5 épocas de paciência                                                          |
| Baralhamento dos dados                      | Sim                                                                            |

## 7 Considerações Finais

Neste trabalho, investigamos como princípios inspirados na cognição humana podem ser aplicados no desenvolvimento de arquiteturas multimodais mais integradas, interpretáveis e eficazes. Partimos da observação de que o cérebro processa e integra informações auditivas e visuais de forma coordenada, com regiões como o *Sulco Temporal Superior* (STS), o *Giro Temporal Superior* (STG) e o Putâmen desempenhando papéis fundamentais na sincronização e fusão sensorial. A pergunta central que motivou esta pesquisa foi: *seria viável reproduzir a habilidade biológica de representação multimodal utilizando como base os modelos e técnicas atuais?*

Como ponto de partida, conduzimos uma fase exploratória utilizando modelos de conversão entre modalidades, treinando *autoencoders* para transformar espectrogramas de áudio em imagens e vice-versa. Esses experimentos demonstraram que era possível encontrar representações latentes compartilhadas entre os dois domínios, mesmo com formatos de entrada e dinâmicas temporais distintas. Esse achado motivou a proposta do módulo Processador AV, incorporado posteriormente na arquitetura principal, com o objetivo de capturar relações estruturais entre modalidades.

A partir dessa base, propusemos a arquitetura X-Net, composta por três módulos principais:

- O **SincAV**, responsável por capturar relações temporais entre áudio e vídeo, modelado como uma tarefa de classificação binária entre amostras sincronizadas e dessincronizadas. O modelo alcançou 81% de acurácia, demonstrando boa capacidade de generalização inclusive em vídeos externos ao conjunto de treinamento.
- O **Processador AV**, que realizou reconstruções unimodais para aprender representações latentes alinhadas entre os domínios. Mesmo sob um volume de dados maior que os modelos de conversão iniciais, o módulo alcançou 71,5% de SSIM e 21,5 dB de PSNR para imagem-imagem, e 62,23% de SSIM e 19,5 dB de PSNR para espectrograma-espectrograma. Além disso, análises de interpolação no espaço latente revelaram transições suaves e semanticamente coerentes entre *embeddings* visuais, sugerindo que o modelo aprendeu um espaço representacional estruturado.
- O **Controlador AV**, cuja função foi sintetizar as informações produzidas pelos dois módulos anteriores em *embeddings* de modulação ( $\mathbf{z}_v, \mathbf{z}_a$ ), utilizadas para condicionar

dinamicamente as *backbones* de vídeo e áudio por meio de *Feature-wise Linear Modulation* (FiLM). A modulação foi aplicada diretamente nas representações finais extraídas por cada *backbone*, antes da concatenação e da *head* de classificação.

Na etapa final de avaliação, a arquitetura completa X-Net foi testada nos conjuntos UCF101 e COIN, ambos amplamente utilizados em tarefas de classificação de ações e atividades humanas. Comparou-se a versão com e sem o *Controlador AV*, isolando o impacto da modulação. Os resultados foram expressivos: no UCF101, a acurácia subiu de 75,5% para 81,7%; no COIN, de 52,3% para 59,1%. Essas melhorias quantitativas validaram a hipótese central do trabalho.

Embora a X-Net não tenha superado os modelos estado da arte, como o MA-LMM ou IMD-B, apresentou um desempenho competitivo no dataset UCF101, aliado a uma arquitetura mais interpretável e analisável. A estrutura modular permitiu o estudo isolado de cada componente, com métricas específicas e visualizações qualitativas, algo frequentemente ausente em abordagens *end-to-end* tradicionais.

## Perspectivas Futuras

A X-Net representou uma contribuição conceitual e experimental relevante para o campo da visão computacional multimodal. Entretanto, sua modularidade e inspiração biológica abrem espaço para uma série de extensões:

- **Modulação hierárquica:** expandir o uso de FiLM para camadas intermediárias dentro das *backbones*, promovendo uma integração ainda mais profunda entre os fluxos;
- **Tarefas autosupervisionadas no Controlador AV:** como predição de modalidade ausente, *contrastive learning* ou identificação de eventos temporais;
- **Robustez multimodal:** testar a arquitetura em cenários com ruído, perda de modalidade ou dados desbalanceados;
- **Implementações leves:** substituir os módulos internos por variantes mais eficientes, visando aplicações em tempo real ou embarcadas;
- **Escalas perceptuais alternativas:** investigar o uso da **escala de Bark** em substituição à escala de Mel para a geração dos espectrogramas de áudio. Por estar mais alinhada à fisiologia da cóclea, a escala de Bark pode oferecer uma representação perceptualmente mais fiel em tarefas que não envolvem diretamente a compreensão de fala (PANDEY et al., 2025; TRIVEDI et al., 2025), permitindo avaliar seu impacto sobre o aprendizado multimodal;

- **Modelagem perceptual inspirada no sistema sensorial humano:** incorporar mecanismos que reproduzam aspectos cognitivos e fisiológicos da visão e audição, com o objetivo de aprimorar a integração multimodal. Entre as possibilidades estão:
  - Representação de **direção e localização de sons**, aproximando a rede de mecanismos de espacialização auditiva;
  - Simulação da **percepção tridimensional**, combinando informações de profundidade e movimento visual;
  - Exploração do **fluxo óptico** e de pistas visuais de movimento como elementos auxiliares para sincronização e correspondência entre modalidades.
- **Amostragem adaptativa de cliques:** explorar diferentes estratégias de segmentação temporal e seleção de quadros, incluindo amostragem densa, aleatória ou guiada por eventos, a fim de avaliar como o padrão de entrada influencia o desempenho do modelo e sua capacidade de generalização.

## Conclusão

Concluimos que a X-Net inaugurou um caminho promissor para o desenvolvimento de modelos neurais multimodais com maior alinhamento biológico, interpretabilidade e flexibilidade. Ao adotar uma abordagem inspirada em como o cérebro opera (por meio de especialização, modulação e integração dinâmica), este trabalho contribuiu para um campo de pesquisa cada vez mais relevante: o da inteligência artificial multimodal com base neurocognitiva.

# Referências

- ABU-EL-HAIJA, S. et al. Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*, 2016.
- ACADEMY, D. S. *Deep Learning Book Brasil*. 2022. Repositório Github. Acessado em: 23 de junho de 2024. Disponível em: <<https://www.deeplearningbook.com.br/>>.
- AFOURAS, T. et al. Self-supervised object detection from audio-visual correspondence. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.: s.n.], 2022. p. 10575–10586.
- ALBERTI, P. W. The anatomy and physiology of the ear and hearing. *Occupational Exposure to Noise: Evaluation, Prevention, and Control*, World Health Organization: Geneva, p. 53–62, 2001.
- ALFASLY, S. et al. Auxiliary audio-textual modalities for better action recognition on vision-specific annotated videos. *Pattern Recognition*, Elsevier, v. 156, p. 110808, 2024.
- APPALARAJU, S. et al. Docformerv2: Local features for document understanding. *arXiv preprint arXiv:2306.01733*, 2023.
- ARANDJELOVIC, R.; ZISSERMAN, A. Look, listen and learn. In: *Proceedings of the IEEE International Conference on Computer Vision*. [S.l.: s.n.], 2017. p. 609–617.
- ARMSTRONG, R.; CUBBIDGE, R. The eye and vision: an overview. *Handbook of Nutrition, Diet and the Eye*, Elsevier, p. 3–9, 2014.
- AYTAR, Y.; VONDRICK, C.; TORRALBA, A. Soundnet: Learning sound representations from unlabeled video. *Advances in Neural Information Processing Systems*, v. 29, 2016.
- AZADI, S. et al. Multi-content gan for few-shot font style transfer. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2018. p. 7564–7573.
- BARBOSA, D. *Conheça os sinais da infeccao de ouvido*. 2018. Repositório Github. Acessado em: 23 de junho de 2024. Disponível em: <<https://fofuuu.com/blog/desatencao-inquietacao-e-coceira-conheca-os-sinais-da-infeccao-de-ouvido/>>.
- BAU, D. et al. Gan dissection: Visualizing and understanding generative adversarial networks. *arXiv preprint arXiv:1811.10597*, 2018.

- BEAUCHAMP, M. S. et al. Integration of auditory and visual information about objects in superior temporal sulcus. *Neuron*, Elsevier, v. 41, n. 5, p. 809–823, 2004.
- BEJIGA, M. B.; MELGANI, F. Gan-based domain adaptation for object classification. In: IEEE. *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*. [S.l.], 2018. p. 1264–1267.
- BÉRARD, A. et al. End-to-end automatic speech translation of audiobooks. In: IEEE. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. [S.l.], 2018. p. 6224–6228.
- BESLE, J. et al. Visual activation and audiovisual interactions in the auditory cortex during speech perception: intracranial recordings in humans. *Journal of Neuroscience*, Soc Neuroscience, v. 28, n. 52, p. 14301–14310, 2008.
- BHATTAD, R. et al. Effects of 2d and 3d max pooling on lstmfeature extraction. In: IEEE. *2023 IEEE 20th India Council International Conference (INDICON)*. [S.l.], 2023. p. 192–196.
- BRITO, G. *Fisiologia Sensorial*. 2017. Repositório Github. Acessado em: 23 de junho de 2024. Disponível em: <<https://resumosmedicina.com.br/tudosobrefisiologiasensorial/>>.
- BROWN, T. B. Language models are few-shot learners. *arXiv preprint ArXiv:2005.14165*, 2020.
- CAMUTO, A. et al. Towards a theoretical understanding of the robustness of variational autoencoders. In: PMLR. *International Conference on Artificial Intelligence and Statistics*. [S.l.], 2021. p. 3565–3573.
- CHE, X. et al. Automatic lecture subtitle generation and how it helps. In: *2017 IEEE 17th International Conference on Advanced Learning Technologies (ICALT)*. [S.l.: s.n.], 2017. p. 34–38.
- DENG, L. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, IEEE, v. 29, n. 6, p. 141–142, 2012.
- DEVLIN, J. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- DÖRFLER, M.; BAMMER, R.; GRILL, T. Inside the spectrogram: Convolutional neural networks in audio processing. In: *2017 International Conference on Sampling Theory and Applications (SampTA)*. [S.l.: s.n.], 2017. p. 152–155.
- DOSOVITSKIY, A. et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.
- DURAK, L.; ARIKAN, O. Short-time fourier transform: two fundamental properties and an optimal implementation. *IEEE*, v. 51, n. 5, p. 1231–1242, 2003.
- EHRET, G.; ROMAND, R. *The central auditory system*. [S.l.]: Oxford University Press, USA, 1997.



- FENG, H.-Z. et al. Shot-vae: semi-supervised deep generative models with label-aware elbo approximations. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. [S.l.: s.n.], 2021. v. 35, n. 8, p. 7413–7421.
- FU, K.-S. Pattern recognition and image processing. *IEEE Transactions on Computers*, IEEE, v. 100, n. 12, p. 1336–1346, 1976.
- GARNER, A. R.; KELLER, G. B. A cortical circuit for audio-visual predictions. *Nature neuroscience*, Nature Publishing Group US New York, v. 25, n. 1, p. 98–105, 2022.
- GELLY, G.; GAUVAIN, J.-L. Optimization of rnn-based speech activity detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, IEEE, v. 26, n. 3, p. 646–656, 2017.
- GIRDHAR, R. et al. Imagebind: One embedding space to bind them all. In: *IEEE (CVPR)*. [S.l.: s.n.], 2023. p. 15180–15190.
- GIRDHAR, R. et al. Imagebind: One embedding space to bind them all. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2023. p. 15180–15190.
- GLOROT, X.; BENGIO, Y. Understanding the difficulty of training deep feedforward neural networks. In: *JMLR Workshop and Conference Proceedings. Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*. [S.l.: s.n.], 2010. p. 249–256.
- GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. *Deep Learning*. 2016. MIT Press. Disponível em: <<https://www.deeplearningbook.org/>>.
- GOODFELLOW, I. J. et al. Generative adversarial nets. *Advances in neural information processing systems*, v. 27, 2014.
- GPIOSENKA, G. *30 Musical Instruments – Image Classification Dataset*. 2022. Repositório Github. Accessed: 2023-11-22. Disponível em: <<https://www.kaggle.com/datasets/gpiosenska/musical-instruments-image-classification>>.
- HASEGAWA, T. et al. Learned audio-visual cross-modal associations in observed piano playing activate the left planum temporale: An fMRI study. *Cognitive Brain Research*, v. 20, n. 3, p. 510–518, 2004.
- HE, B. et al. Ma-lmm: Memory-augmented large multimodal model for long-term video understanding. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2024. p. 13504–13514.
- HE, K. et al. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: *Proceedings of the IEEE International Conference on Computer Vision*. [S.l.: s.n.], 2015. p. 1026–1034.
- HORI, C. et al. Attention-based multimodal fusion for video description. In: *Proceedings of the IEEE International Conference on Computer Vision*. [S.l.: s.n.], 2017. p. 4193–4202.
- HUANG, D.-A. et al. What makes a video a video: Analyzing temporal information in video understanding models and datasets. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2018. p. 7366–7375.

- HUANG, Y. et al. Layoutlmv3: Pre-training for document ai with unified text and image masking. In: *Proceedings of the 30th ACM International Conference on Multimedia*. New York, NY, USA: Association for Computing Machinery, 2022. (MM '22), p. 4083–4091. ISBN 9781450392037. Disponível em: <<https://doi.org/10.1145/3503161.3548112>>.
- IOFFE, S.; SZEGEDY, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: PMLR. *International Conference on Machine Learning*. [S.l.], 2015. p. 448–456.
- ISLAM, M. M. et al. Efficient movie scene detection using state-space transformers. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2023. p. 18749–18758.
- JACKSON, Z. *Free Spoken Digit Dataset (FSDD)*. 2016. Repositório Github. Acessado em: 2023-11-22. Disponível em: <<https://github.com/Jakobovski/free-spoken-digit-dataset>>.
- JAISWAL, A. et al. A survey on contrastive self-supervised learning. *Technologies*, MDPI, v. 9, n. 1, p. 2, 2020.
- JI, S. et al. 3d convolutional neural networks for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, IEEE, v. 35, n. 1, p. 221–231, 2012.
- JIANG, Z.; SWEETSER, P. Gan-assisted yuv pixel art generation. In: SPRINGER. *Australasian Joint Conference on Artificial Intelligence*. [S.l.], 2022. p. 595–606.
- JORDON, J.; YOON, J.; SCHAAR, M. V. D. Pate-gan: Generating synthetic data with differential privacy guarantees. In: *International Conference on Learning Representations*. [S.l.: s.n.], 2018.
- KASHEVNIK, A. et al. Multimodal corpus design for audio-visual speech recognition in vehicle cabin. *IEEE Access*, v. 9, p. 34986–35003, 2021.
- KHURANA, A. et al. Tri-integrated convolutional neural network for audio image classification using mel-frequency spectrograms. *Multimedia Tools and Applications*, Springer, v. 82, n. 4, p. 5521–5546, 2023.
- KINGMA, D. P.; WELLING, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.
- KNÖPFEL, T. et al. Audio-visual experience strengthens multisensory assemblies in adult mouse visual cortex. *Nature Communications*, Nature Publishing Group UK London, v. 10, n. 1, p. 5684, 2019.
- KORBAR, B.; TRAN, D.; TORRESANI, L. Cooperative learning of audio and video models from self-supervised synchronization. *Advances in Neural Information Processing Systems*, v. 31, 2018.
- KRISHNAN, R.; RAJPURKAR, P.; TOPOL, E. J. Self-supervised learning in medicine and healthcare. *Nature Biomedical Engineering*, Nature Publishing Group UK London, v. 6, n. 12, p. 1346–1352, 2022.

- KRIZHEVSKY, A.; SUTSKEVER, I.; HINTON, G. E. Imagenet classification with deep convolutional neural networks. In: PEREIRA, F. et al. (Ed.). *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2012. v. 25. Disponível em: <[https://proceedings.neurips.cc/paper/\\_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf](https://proceedings.neurips.cc/paper/_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf)>.
- LI, H. et al. Hierarchical feature extraction with local neural response for image recognition. *IEEE Transactions on Cybernetics*, IEEE, v. 43, n. 2, p. 412–424, 2013.
- LI, K. et al. Videomamba: State space model for efficient video understanding. In: SPRINGER. *European Conference on Computer Vision*. [S.l.], 2024. p. 237–255.
- LI, Q. et al. Medical image classification with convolutional neural network. In: IEEE. *2014 13th International Conference on Control Automation Robotics & Vision (ICARCV)*. [S.l.], 2014. p. 844–848.
- LIN, Y.-B.; WANG, Y.-C. F. Audiovisual transformer with instance attention for audio-visual event localization. In: *Proceedings of the Asian Conference on Computer Vision (ACCV)*. [s.n.], 2020. Disponível em: <[https://openaccess.thecvf.com/content/ACCV2020/papers/Lin/\\_Audiovisual/\\_Transformer/\\_with/\\_Instance/\\_Attention/\\_for/\\_Audio-Visual/\\_Event/\\_Localization/\\_ACCV/\\_2020/\\_paper.pdf](https://openaccess.thecvf.com/content/ACCV2020/papers/Lin/_Audiovisual/_Transformer/_with/_Instance/_Attention/_for/_Audio-Visual/_Event/_Localization/_ACCV/_2020/_paper.pdf)>.
- MAHMUD, S. S. et al. Vader: A variational autoencoder to infer planetary masses and gas-dust disk properties around young stars. *arXiv e-prints*, p. arXiv-2509, 2025.
- MALMIERCA, M. S. Auditory system. In: *The Rat Nervous System*. [S.l.]: Elsevier, 2015. p. 865–946.
- MATHWORKS. *Train a Variational Autoencoder (VAE) to Generate Images*. 2024. Repositório Github. Acessado em: 23 de junho de 2024. Disponível em: <<https://www.mathworks.com/help/deeplearning/ug/train-a-variational-autoencoder-vae-to-generate-images.html>>.
- MCGURK, H.; MACDONALD, J. Hearing lips and seeing voices. *Nature*, Nature Publishing Group UK London, v. 264, n. 5588, p. 746–748, 1976.
- MISRA, I.; MAATEN, L. v. d. Self-supervised learning of pretext-invariant representations. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2020. p. 6707–6717.
- MOCANU, B.; TAPU, R. Automatic subtitle synchronization and positioning system dedicated to deaf and hearing impaired people. *IEEE Access*, v. 9, p. 139544–139555, 2021.
- MOESKOPS, P. et al. Automatic segmentation of mr brain images with a convolutional neural network. *IEEE Transactions on Medical Imaging*, IEEE, v. 35, n. 5, p. 1252–1261, 2016.
- MOHANTY, S. P. *Musical Instrument's Sound Dataset*. 2022. Repositório Github. Accessed: 2023-11-22. Disponível em: <<https://www.kaggle.com/datasets/soumendraprasad/musical-instruments-sound-dataset>>.

- MORRISON, C. et al. Collaborating with computer vision systems: An exploration of audio feedback. In: *Proceedings of the 2014 Conference on Designing Interactive Systems*. New York, NY, USA: Association for Computing Machinery, 2014. (DIS '14), p. 229–238. ISBN 9781450329026. Disponível em: <<https://doi.org/10.1145/2598510.2598519>>.
- NADKARNI, P. M.; OHNO-MACHADO, L.; CHAPMAN, W. W. Natural language processing: an introduction. *Journal of the American Medical Informatics Association*, BMJ Group BMA House, Tavistock Square, London, WC1H 9JR, v. 18, n. 5, p. 544–551, 2011.
- NILSSON, J.; AKENINE-MÖLLER, T. Understanding ssim. *arXiv preprint arXiv:2006.13846*, 2020.
- NING, H. et al. Audio description from image by modal translation network. *Neurocomputing*, Elsevier, v. 423, p. 124–134, 2021.
- NODA, K. et al. Audio-visual speech recognition using deep learning. *Applied Intelligence*, Springer, v. 42, p. 722–737, 2015.
- OWENS, A.; EFROS, A. A. Audio-visual scene analysis with self-supervised multisensory features. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. [S.l.: s.n.], 2018. p. 631–648.
- PANDEY, A. et al. Multilayer perceptron classification for multilingual speech detection. *Procedia Computer Science*, Elsevier, v. 260, p. 447–456, 2025.
- PARTHASARATHY, S.; SUNDARAM, S. Detecting expressions with multimodal transformers. In: IEEE. *2021 IEEE Spoken Language Technology Workshop (SLT)*. [S.l.], 2021. p. 636–643.
- PEREZ, E. et al. Film: Visual reasoning with a general conditioning layer. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. [S.l.: s.n.], 2018. v. 32, n. 1.
- RAMANISHKA, V. et al. Multimodal video description. In: *Proceedings of the 24th ACM International Conference on Multimedia*. New York, NY, USA: Association for Computing Machinery, 2016. (MM '16), p. 1092–1096. ISBN 9781450336031. Disponível em: <<https://doi.org/10.1145/2964284.2984066>>.
- RANKIN, S. K. *Classification of Meows and Woofs: Part 1 Spectrograms*. 2017. Repositório Github. Acessado em: 23 de junho de 2024. Disponível em: <<https://www.summerrankin.com/dogandponyshow/2017/10/16/catdog>>.
- SCHEEF, L. et al. Multimodal motion processing in area v5/mt: evidence from an artificial class of audio-visual events. *Brain Research*, Elsevier, v. 1252, p. 94–104, 2009.
- SCHILLER, P. H. The central visual system. *Vision Research*, Elsevier, v. 26, n. 9, p. 1351–1386, 1986.
- ŠEVO, I.; AVRAMOVIĆ, A. Convolutional neural network based automatic object detection on aerial images. *IEEE Geoscience and Remote Sensing Letters*, IEEE, v. 13, n. 5, p. 740–744, 2016.

- SOOMRO, K.; ZAMIR, A. R.; SHAH, M. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.
- STERLING, A. et al. Isnn: Impact sound neural network for audio-visual object classification. In: FERRARI, V. et al. (Ed.). *Computer Vision – ECCV 2018*. Cham: Springer International Publishing, 2018. p. 578–595. ISBN 978-3-030-01267-0.
- STURMEL, N.; DAUDET, L. et al. Signal reconstruction from stft magnitude: A state of the art. In: *International Conference on Digital Audio Effects (DAFx)*. [S.l.: s.n.], 2011. p. 375–386.
- SUMMAIRA, J. et al. Recent advances and trends in multimodal deep learning: A review. *CoRR*, abs/2105.11087, 2021.
- TANG DAJUN DING, Y. R. Y. Coin: A large-scale dataset for comprehensive instructional video analysis. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.: s.n.], 2019.
- TAYLOR, C. The multimodal approach in audiovisual translation. *Target. International Journal of Translation Studies*, John Benjamins, v. 28, n. 2, p. 222–236, 2016. ISSN 0924-1884. Disponível em: <<https://www.jbe-platform.com/content/journals/10.1075/target.28.2.04tay>>.
- TIAN, H. et al. Multimodal deep representation learning for video classification. *World Wide Web*, Springer, v. 22, n. 3, p. 1325–1341, 2019.
- TOLSTIKHIN, I. et al. Wasserstein auto-encoders. *arXiv preprint arXiv:1711.01558*, 2017.
- TONG, Z. et al. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in Neural Information Processing Systems*, v. 35, p. 10078–10093, 2022.
- TOVÉE, M. J. *An introduction to the visual system*. [S.l.]: Cambridge University Press, 2008.
- TRAN, D. et al. Learning spatiotemporal features with 3d convolutional networks. In: *Proceedings of the IEEE international conference on computer vision*. [S.l.: s.n.], 2015. p. 4489–4497.
- TRAN, L. D.; NGUYEN, S. M.; ARAI, M. Gan-based noise model for denoising real images. In: *Proceedings of the Asian Conference on Computer Vision*. [S.l.: s.n.], 2020.
- TRIVEDI, N. et al. Cosshi: Llm-powered large-scale multitask hinglish dataset for speech forensics. In: IEEE. *2025 International Conference on Asian Language Processing (IALP)*. [S.l.], 2025. p. 135–140.
- VASWANI, A. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.
- VIDHYA, A. *Generate Your Own Dataset using GAN*. 2021. Repositório Github. Acessado em: 23 de junho de 2024. Disponível em: <<https://www.analyticsvidhya.com/blog/2021/04/generate-your-own-dataset-using-gan/>>.

- VIDHYA, A. *Image Processing Using CNN: A Beginner's Guide*. 2021. Repositório Github. Accessed: 23 de junho de 2024. Disponível em: <<https://www.analyticsvidhya.com/blog/2021/06/image-processing-using-cnn-a-beginners-guide/>>.
- VINCENT, P. et al. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research*, v. 11, n. 12, 2010.
- WANG, J.; JIN, L.; DING, K. Lilt: A simple yet effective language-independent layout transformer for structured document understanding. *arXiv preprint arXiv:2202.13669*, 2022.
- XIAO, F. et al. Audiovisual slowfast networks for video recognition. *arXiv preprint arXiv:2001.08740*, 2020.
- XU, D. et al. Self-supervised spatiotemporal learning via video clip order prediction. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2019. p. 10334–10343.
- YANG, K. et al. 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 (CVPR)*. [S.l.: s.n.], 2022. p. 8227–8237.
- YANG, P.-T.; SU, F.-G.; WANG, Y.-C. F. Diverse audio-to-image generation via semantics and feature consistency. In: IEEE. *2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. [S.l.], 2020. p. 1188–1192.
- YOUNESI, A. et al. A comprehensive survey of convolutions in deep learning: Applications, challenges, and future trends. *IEEE Access*, IEEE, v. 12, p. 41180–41218, 2024.
- ŻELASZCZYK, M.; MAŃDZIUK, J. Audio-to-image cross-modal generation. In: IEEE. *2022 International Joint Conference on Neural Networks (IJCNN)*. [S.l.], 2022. p. 1–8.

**DADOS CURRICULARES**

| <b>IDENTIFICAÇÃO</b>                                                                                                                                                                                        |                                                                                                                   |
|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------|
|                                                                                                                                                                                                             | Gustavo Furlanetto<br>12 de outubro de 2000                                                                       |
| <b>Nacionalidade</b>                                                                                                                                                                                        | Brasileiro                                                                                                        |
| <b>Nome em citações bibliográficas</b>                                                                                                                                                                      | Furlanetto, Gustavo<br>Furlanetto, G.                                                                             |
| <b>Currículo Lattes</b>                                                                                                                                                                                     | <a href="http://lattes.cnpq.br/2537831061483890">http://lattes.cnpq.br/2537831061483890</a>                       |
| <b>ORCID</b>                                                                                                                                                                                                | <a href="https://orcid.org/0009-0006-3687-123X">https://orcid.org/0009-0006-3687-123X</a>                         |
| <b>FORMAÇÃO ACADÊMICA</b>                                                                                                                                                                                   |                                                                                                                   |
| <b>2019/2022</b>                                                                                                                                                                                            | Bacharelado em Ciências da Computação<br>UNESP - Instituto de Geociências e Ciências Exatas - Câmpus de Rio Claro |
| <b>PRODUÇÃO BIBLIOGRÁFICA</b>                                                                                                                                                                               |                                                                                                                   |
|                                                                                                                                                                                                             |                                                                                                                   |
| <b>PARTICIPAÇÃO EM BANCAS E ORIENTAÇÕES</b>                                                                                                                                                                 |                                                                                                                   |
| <b>Bancas de trabalhos de conclusão</b>                                                                                                                                                                     |                                                                                                                   |
|                                                                                                                                                                                                             |                                                                                                                   |
| <b>Orientações</b>                                                                                                                                                                                          |                                                                                                                   |
|                                                                                                                                                                                                             |                                                                                                                   |
| <b>PARTICIPAÇÃO EM EVENTOS CIENTÍFICOS</b>                                                                                                                                                                  |                                                                                                                   |
| XII WORKSHOP DO PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO DA UNESP (WPPGCC 2023), 11., 2023, Remoto. Modelos multimodais: Explorando a união entre áudio e imagem em dados Multimedia (Seminário). |                                                                                                                   |
| XIII WORKSHOP DO PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO DA UNESP (WPPGCC 2024), 11., 2024, Remoto. Convergência Multimodal para Processamento Audiovisual de Dados Multimedia (Seminário).      |                                                                                                                   |