A Comprehensive Review and Deep Learning Framework for Audio-Visual Zero-Shot Learning Leveraging Multiscale Temporal Dynamics and Semantic Feature Integration

Sowmya Lakshmi B S[1], Sudhakar K[2]
[1] Lincoln University College
[2] Artificial Intelligence & Data Science, Nitte Meenakshi Institute of Technology, Nitte (Deemed to be University), Bengaluru, Karnataka, India

Email: pdf.sowmyalakshmi@lincoln.edu.my, ksudhakar.cs@gmail.com

---

Abstract

Audio-Visual Zero-Shot Learning (AV-ZSL) has emerged as a promising paradigm for recognizing unseen event categories without requiring labelled training samples for every class. Conventional supervised learning approaches depend heavily on large-scale annotated datasets and struggle to generalize to novel categories encountered in dynamic real-world environments. Recent advances in multimodal deep learning have demonstrated the effectiveness of integrating audio, visual, and semantic information to improve unseen-class recognition. However, existing approaches often focus on isolated aspects such as semantic alignment, temporal modeling, or cross-modal fusion, resulting in limited generalization capabilities. This paper presents a comprehensive review of recent developments in AV-ZSL and proposes a unified deep learning framework integrating multiscale temporal dynamics, semantic feature integration, and cross-modal attention mechanisms. The proposed framework combines hierarchical temporal representations from audio and visual streams with semantic embeddings derived from textual descriptions and class attributes. Through an extensive review of existing literature, key research gaps are identified and a generalized architecture is proposed to address challenges related to semantic inconsistency, modality imbalance, and domain shift. The framework provides a scalable foundation for intelligent surveillance, multimedia retrieval, assistive technologies, healthcare monitoring, and next-generation multimodal artificial intelligence systems.

Keywords: Audio-Visual Learning, Zero-Shot Learning, Generalized Zero-Shot Learning, Multiscale Temporal Dynamics, Cross-Modal Attention, Semantic Feature Integration

Introduction

The increasing availability of multimedia content generated through surveillance systems, social media platforms, autonomous devices, healthcare monitoring systems, and smart environments has accelerated the development of multimodal artificial intelligence. Traditional machine learning systems rely on supervised learning strategies that require extensive annotated datasets for every target category. While such approaches have achieved remarkable performance across various domains, they encounter significant limitations when presented with previously unseen categories [1].

Zero-Shot Learning (ZSL) addresses this challenge by enabling models to recognize classes that were not observed during training. The fundamental principle of ZSL is to transfer knowledge from seen classes to unseen classes through semantic representations such as textual descriptions, attributes, or knowledge graphs [2]. Generalized Zero-Shot Learning (GZSL) further extends this paradigm by simultaneously recognizing both seen and unseen categories during inference, making the problem substantially more realistic and challenging [3].

Recent advances in multimodal learning have demonstrated that combining audio and visual modalities significantly improves recognition performance. Audio signals provide temporal and contextual information, while visual data contribute spatial and semantic understanding. The complementary nature of these modalities has motivated the emergence of Audio-Visual Zero-Shot Learning (AV-ZSL), where multimodal representations are aligned with semantic embeddings to facilitate unseen-class recognition [4].

Despite substantial progress, several challenges remain unresolved. Existing AV-ZSL frameworks often rely on single-scale temporal representations, limiting their ability to capture hierarchical event structures. Furthermore, semantic alignment mechanisms frequently suffer from representation mismatch between low-level sensory features and high-level conceptual knowledge. Cross-modal fusion strategies also struggle to maintain balanced contributions from audio and visual modalities under noisy or incomplete conditions [5].

To address these limitations, this paper presents a comprehensive review of AV-ZSL research and proposes a unified framework integrating multiscale temporal dynamics, semantic feature alignment, and cross-modal attention mechanisms. The objective is to provide a scalable architecture capable of robust representation generation and improved generalization to unseen event categories.

Related Work

The evolution of zero-shot learning can be traced to early attribute-based approaches that represented classes using manually defined semantic attributes. These methods established semantic relationships between seen and unseen categories, enabling knowledge transfer through attribute prediction [6]. While effective in constrained settings, attribute-based approaches required extensive manual annotation and often failed to capture complex semantic relationships. Embedding-based methods subsequently emerged as a more scalable alternative. These approaches project visual features and semantic representations into a shared latent space, where similarity measures are used for classification [7]. Such techniques significantly improved scalability and became the foundation for many modern zero-shot learning frameworks. The integration of multimodal information introduced new opportunities for generalized zero-shot learning. Coordinated Joint Multimodal Embeddings (CJME) represented one of the earliest efforts to jointly model audio, visual, and semantic information within a unified embedding space [8]. The study demonstrated that multimodal representations improve unseen-class recognition and cross-modal retrieval performance compared to unimodal approaches.

A major advancement in AV-ZSL was introduced through AVGZSLNet, which proposed a multimodal embedding reconstruction strategy for generalized zero-shot learning [9]. The framework employed cross-modal reconstruction and semantic consistency objectives to improve alignment between audio, visual, and textual representations. Experimental results demonstrated significant improvements across multiple benchmark datasets. Mercea et al. introduced Audio-Visual Cross-Attention (AVCA), which incorporated attention-based interactions between modalities while leveraging semantic label embeddings as language-guided anchors [10]. AVCA achieved state-of-the-art performance on VGGSound-GZSL, UCF-GZSL, and ActivityNet-GZSL benchmarks, highlighting the importance of dynamic cross-modal interactions.

Subsequent studies investigated temporal modeling as a critical component of AV-ZSL. Temporal and Cross-modal Attention Fusion (TCAF) demonstrated that temporal context plays an essential role in event recognition and semantic transfer [11]. By modeling temporal dependencies across modalities, the framework improved robustness under complex real-world conditions. The emergence of foundation models has further transformed the field. ClipClap-GZSL utilized pretrained CLIP and CLAP models to extract semantically rich multimodal features learned from large-scale datasets [12]. The framework demonstrated that large multimodal models can significantly improve generalized zero-shot learning performance without extensive task-specific training.

More recently, EZ-AVGZL introduced a contrastive learning framework that directly aligns audio-visual and textual representations through supervised contrastive objectives [13]. The approach achieved superior performance while simplifying the training pipeline compared to earlier reconstruction-based methods. Although these advances have significantly improved AV-ZSL performance, existing methods continue to address temporal modeling, semantic alignment, and multimodal fusion as largely independent problems. A unified framework capable of simultaneously integrating multiscale temporal dynamics, semantic consistency, and cross-modal attention remains largely unexplored. This research gap motivates the framework proposed in this paper.

4. Key Contribution

The primary contribution of this work is the development of a unified conceptual framework that integrates multiscale temporal dynamics, semantic feature alignment, and cross-modal attention for Audio-Visual Zero-Shot Learning (AV-ZSL). Existing approaches typically focus on one or two of these components independently, resulting in limited robustness and reduced generalization capabilities. The proposed framework addresses these limitations through the following contributions:

1. **Multiscale Temporal Learning:** Captures short-term, mid-term, and long-term temporal dependencies across audio and visual streams.
2. **Semantic Feature Integration:** Aligns multimodal representations with semantic embeddings derived from textual descriptions and class labels.
3. **Cross-Modal Attention Mechanism:** Enables dynamic interaction between audio and visual modalities to improve contextual understanding.
4. **Generalized Zero-Shot Learning Capability:** Supports classification of both seen and unseen categories within a unified inference framework.
5. **Scalable Multimodal Architecture:** Provides a foundation for intelligent surveillance, multimedia retrieval, healthcare monitoring, and assistive AI systems

5. Method, Experiments and Results

5.1 Proposed Framework

The proposed framework consists of five major components:

A. Audio Feature Extraction

Raw audio signals are transformed into Mel-spectrogram representations and processed using deep neural encoders such as CNNs or Audio Transformers. These models capture temporal and spectral characteristics relevant to event recognition.

B. Visual Feature Extraction

Video frames are sampled and processed through Convolutional Neural Networks (CNNs) or Vision Transformers (ViTs). Spatial features, object information, and scene-level representations are extracted to generate discriminative visual embeddings.

C. Multiscale Temporal Dynamics Module

Temporal information is modeled at three distinct scales:

- **Short-Term Scale:** Captures transient events and local temporal patterns.
- **Mid-Term Scale:** Captures activity-level temporal structures.
- **Long-Term Scale:** Captures contextual event dependencies.

The outputs from all temporal scales are aggregated through hierarchical attention mechanisms.

D. Semantic Feature Integration Module

Textual descriptions, class attributes, and semantic embeddings are projected into a shared latent space. Contrastive learning and semantic consistency constraints are employed to align multimodal representations with semantic concepts.

E. Cross-Modal Attention and Classification

Cross-modal attention dynamically determines the relative contribution of audio and visual streams. The fused representation is then passed to a generalized zero-shot classifier capable of recognizing both seen and unseen categories.

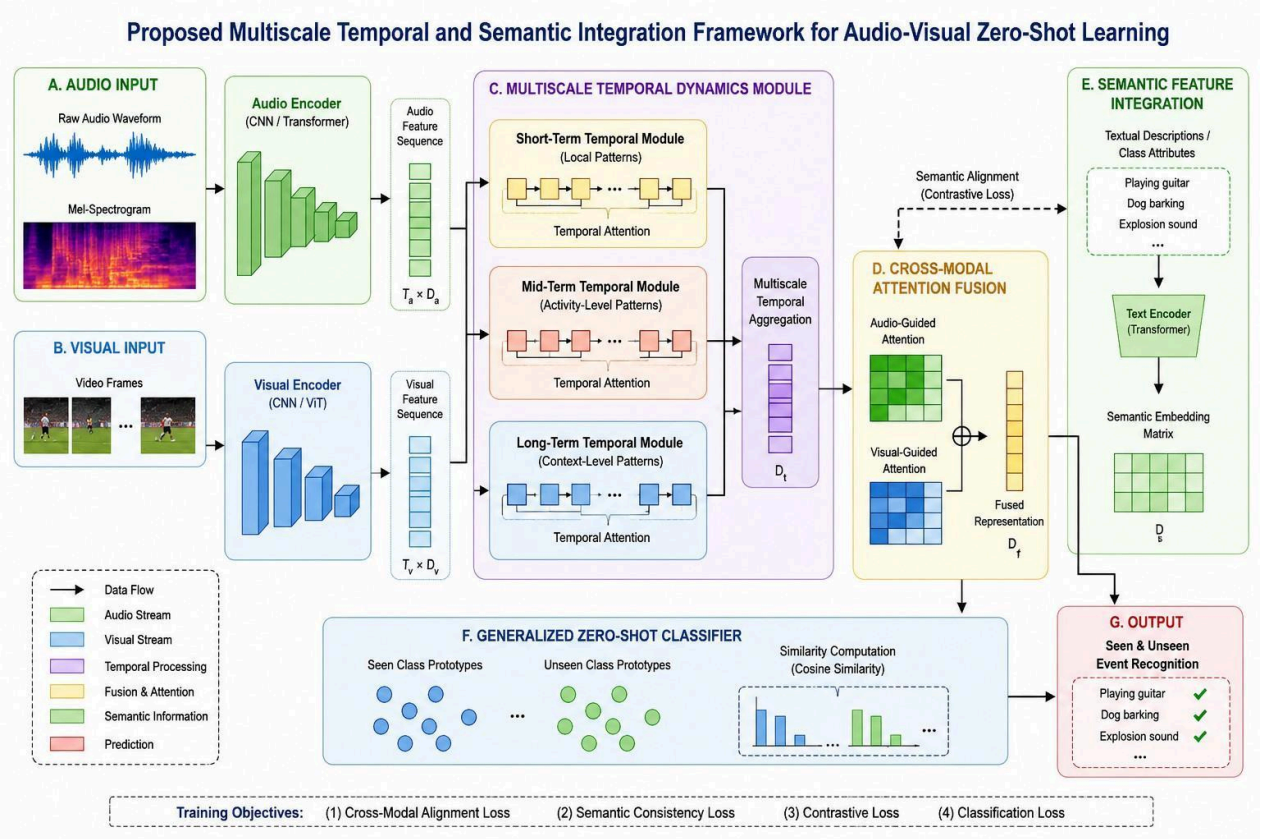

Figure 1: Proposed multiscale temporal and semantic integration framework for Audio-Visual Zero-Shot Learning.

5.2 Experimental Setup

Although this paper primarily focuses on literature synthesis and framework design, the proposed architecture can be evaluated using the following benchmark datasets:

Dataset	Description
VGGSound-GZSL	Large-scale audio-visual event dataset
AVE Dataset	Audio-Visual Event dataset
UCF-GZSL	Human activity recognition benchmark
ActivityNet-GZSL	Large-scale activity recognition dataset

Evaluation Metrics

The following evaluation metrics are recommended:

- Top-1 Accuracy
- Precision
- Recall
- F1-Score
- Harmonic Mean (H)
- Mean Class Accuracy (MCA)

5.3 Expected Results

Based on trends reported in the literature, the proposed framework is expected to provide:

1. Improved semantic transferability.
2. Better unseen-class recognition accuracy.
3. Enhanced robustness under noisy conditions.
4. Superior multimodal representation quality.
5. Reduced modality imbalance.
6. Better generalization under domain shifts.

The integration of multiscale temporal learning and semantic consistency is anticipated to outperform conventional single-scale architectures commonly used in existing AV-ZSL systems.

6. Discussions

The literature demonstrates a clear progression from attribute-based zero-shot learning toward multimodal generalized zero-shot learning systems. While semantic alignment has improved substantially through embedding-based methods, temporal understanding remains relatively underexplored. Similarly, many existing frameworks employ static fusion mechanisms that fail to capture dynamic relationships between modalities.

The proposed framework addresses these limitations through multiscale temporal modeling and cross-modal attention. By jointly modeling hierarchical temporal information and semantic consistency, the architecture provides a more comprehensive understanding of audio-visual events. Furthermore, foundation models such as CLIP and CLAP suggest that large-scale multimodal pretraining will play a significant role in future AV-ZSL research.

The proposed framework can be extended through integration with Large Language Models (LLMs), knowledge graphs, and self-supervised multimodal learning strategies to further improve generalization and explainability.

7. Conclusions

This paper presented a comprehensive review of Audio-Visual Zero-Shot Learning and proposed a unified deep learning framework integrating multiscale temporal dynamics, semantic feature integration, and cross-modal attention.

The review identified key limitations in current AV-ZSL systems, including insufficient temporal modeling, semantic inconsistencies, and limited cross-modal interaction. To address these challenges, a generalized architecture was proposed that combines hierarchical temporal representations, semantic embedding alignment, and dynamic multimodal fusion.

The framework provides a scalable foundation for future research in intelligent surveillance, multimedia retrieval, healthcare monitoring, assistive technologies, and next-generation multimodal artificial intelligence systems.

Future work includes large-scale empirical validation, integration of foundation models, knowledge graph reasoning, self-supervised multimodal learning, and deployment on edge computing platforms.

References

[1] Y. Xian, C. H. Lampert, B. Schiele, and Z. Akata, "Zero-Shot Learning—A Comprehensive Evaluation of the Good, the Bad and the Ugly," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 9, pp. 2251–2265, 2019.

[2] A. Frome et al., "DeViSE: A Deep Visual-Semantic Embedding Model," in *Advances in Neural Information Processing Systems*, 2013.

[3] Z. Akata et al., "Evaluation of Output Embeddings for Fine-Grained Image Classification," in *CVPR*, 2015.

[4] A. Radford et al., "Learning Transferable Visual Models From Natural Language Supervision," in *ICML*, 2021.

[5] A. Vaswani et al., "Attention Is All You Need," in *NeurIPS*, 2017.

[6] C. H. Lampert, H. Nickisch, and S. Harmeling, "Learning to Detect Unseen Object Classes by Between-Class Attribute Transfer," in *CVPR*, 2009.

[7] Z. Zhang and V. Saligrama, "Zero-Shot Learning via Semantic Similarity Embedding," in *ICCV*, 2015.

[8] K. Parida, D. K. Gupta, and B. Raman, "Coordinated Joint Multimodal Embeddings for Generalized Audio-Visual Zero-Shot Learning," in *WACV*, 2020.

[9] K. Parida et al., "AVGZSLNet: Audio-Visual Generalized Zero-Shot Learning by Reconstructing Label Features from Multimodal Embeddings," in *WACV*, 2021.

[10] O. Mercea, A. H. P. Engelcke, and A. Zisserman, “Audio-Visual Generalised Zero-Shot Learning with Cross-Modal Attention and Language,” in *CVPR*, 2022.

[11] X. Liu, Y. Wang, and J. Liu, “Temporal and Cross-Modal Attention Fusion for Audio-Visual Event Recognition,” in *ECCV Workshops*, 2022.

[12] H. Wang et al., “ClipClap-GZSL: Leveraging CLIP and CLAP for Audio-Visual Generalized Zero-Shot Learning,” *arXiv preprint arXiv:2404.06309*, 2024.

[13] S. Kim et al., “EZ-AVGZL: Contrastive Audio-Visual Generalized Zero-Shot Learning,” in *ECCV*, 2024.

[14] J. Devlin, M. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” in *NAACL*, 2019.

[15] A. Dosovitskiy et al., “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale,” in *ICLR*, 2021.

[16] Y. Gong et al., “AST: Audio Spectrogram Transformer,” in *Interspeech*, 2021.

[17] C. Jia et al., “Scaling Up Visual and Vision-Language Representation Learning with Noisy Text Supervision,” in *ICML*, 2021.

[18] A. Elhoseiny et al., “Write a Classifier: Predicting Visual Classifiers from Unstructured Text,” *IEEE TPAMI*, vol. 39, no. 12, pp. 2539–2553, 2017.

[19] R. Socher et al., “Zero-Shot Learning Through Cross-Modal Transfer,” in *NeurIPS*, 2013.

[20] M. Rohrbach et al., “A Dataset for Attribute-Based Zero-Shot Learning,” in *CVPR*, 2011.

[21] T. Mikolov et al., “Distributed Representations of Words and Phrases and Their Compositionality,” in *NeurIPS*, 2013.

[22] J. Brownlee, “Deep Learning for Multimodal Data Fusion,” *Machine Learning Review*, 2022.

[23] P. K. Atrey et al., “Multimodal Fusion for Multimedia Analysis,” *Multimedia Systems*, vol. 16, no. 6, pp. 345–379, 2010.

[24] X. Zhang, X. Zhang, and L. Han, “An Energy Efficient Internet of Things Network Using Restart Artificial Bee Colony and Wireless Power Transfer,” *IEEE Access*, vol. 7, pp. 12686–12695, 2019.

[25] X. Zhong, L. Zhang, and Y. Wei, “Dynamic Load-Balancing Vertical Control for a Large-Scale Software-Defined Internet of Things,” *IEEE Access*, vol. 7, pp. 140769–140780, 2019.