Journal Homepage: -www.journalijar.com

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH (IJAR)

Article DOI:10.21474/IJAR01/24529
DOI URL: http://dx.doi.org/10.21474/IJAR01/24529

RESEARCH ARTICLE

ASYNCHRONOUS, SELF-SUPERVISED AND UNCERTAINTY-AWARE AUDIO-VISUAL FUSION FOR VECTOR SURVEILLANCE IN RESOURCE-CONSTRAINED ENVIRONMENTS

Serigne Modou Kara Samb[1], Okangondo Loshima Junior[2], Thierno Birame Sambe[3], Idy Diop[4] and Moussa Dethie Sarr[1]

1. UFR of Science and Technology, Iba Der Thiam University, Thiès, Senegal.
2. Faculty of Science and Technology, Cheikh Anta Diop University of Dakar, Dakar, Senegal.
3. Department of Mathematics, Institut Supérieur d'Informatique (ISI), Dakar, Senegal.
4. École Supérieure Polytechnique, Cheikh Anta Diop University of Dakar, Dakar, Senegal.

Manuscript Info

Manuscript History
Received: 08 September 2026
Final Accepted: 10 October 2026
Published: November 2026

Key words:-
self-supervised learning; audio-visual fusion; uncertainty estimation; vector surveillance; low-resource environments

Abstract

Background: Automated surveillance of disease vectors is a major public health need in low-resource settings, where annotated data are scarce, low-cost sensors produce noisy signals and acoustic and visual streams are rarely synchronised. Supervised multimodal approaches assume abundant labels and precise temporal alignment, two conditions seldom met in the field.

Methods: We propose a self-supervised audio-visual fusion framework that tolerates temporal asynchrony and accounts for sensor reliability. A ResNet-18 audio encoder processes Mel spectrograms of mosquito wingbeats and a MobileNetV2 visual encoder processes RGB images. Both are trained without labels using an InfoNCE contrastive objective in which every visual segment falling within a temporal window of an audio segment is treated as a positive match. An uncertainty module attached to each branch drives an adaptive fusion that down-weights the less reliable modality. Representations were evaluated by linear probing against unimodal and supervised multimodal baselines, and under acoustic noise, low light and audio-visual time shifts.

Results: After 100 training epochs, the mean cosine similarity reached 0.82 for positive pairs and 0.21 for negative pairs. The proposed approach achieved 92.8% accuracy (F1-score 0.92), compared with 89.7% (0.89) for the supervised multimodal baseline, 86.1% for the visual-only model and 84.3% for the audio-only model.

"© 2026 by the Author(s). Published by IJAR under CC BY 4.0. Unrestricted use allowed with credit to the author."

At a signal-to-noise ratio of 5 dB, performance dropped by 4% against 12% for the audio-only model. With random shifts of up to ±500 ms between streams, the loss remained below 3%, whereas strict-alignment approaches lost up to 10%. Inference took 42 ms per sample on a standard four-core CPU.

Corresponding Author:- Serigne Modou Kara Samb
Address:-UFR of Science and Technology, Iba Der Thiam University, Thiès, Senegal.

Conclusion: Combining self-supervised contrastive learning, tolerance to asynchrony and uncertainty-weighted fusion yields robust multimodal representations while reducing dependence on manual annotation. The framework runs on low-cost hardware and is suited to operational vector surveillance in resource-constrained environments.

Introduction:-

Vector-borne diseases remain a major public health challenge, particularly in low-resource settings where surveillance capacity is limited. Malaria alone caused an estimated 263 million cases and 597,000 deaths in 2023, 95% of them in the WHO African Region [1]. Vector detection and monitoring still rely largely on manual methods that are costly and depend on specialised entomological expertise, which hinders large-scale deployment and continuity of operations [2]. Artificial intelligence and deep learning are increasingly used to automate vector surveillance from environmental, acoustic and visual data, with the aim of extending its spatial and temporal coverage where human and material resources are scarce [3, 4].

A first line of work relies on visual data. Segmentation and classification models applied to high-resolution or satellite imagery map vector habitats and assess epidemiological risk [3], and deep neural networks classify mosquito species from images of their bodies and wings [5, 6, 2]. A second line exploits acoustic signatures: wingbeat sounds allow species classification and activity detection, even in the presence of ambient noise [7, 8], and large annotated acoustic datasets recorded in the field are now available [9]. Although promising, these unimodal approaches remain sensitive to acquisition conditions and to the variable quality of sensors, which limits their robustness in the field.

Multimodal approaches have been proposed to improve robustness by combining complementary sources of information [10]. Audio-visual and multi-sensor fusion frameworks outperform unimodal methods by exploiting the complementarity of modalities, including in low-resource contexts [11, 12], and the fusion of multispectral and radar imagery shows strong potential for habitat mapping [13, 14]. Most of these approaches, however, are supervised: they require large annotated datasets and assume precise temporal synchronisation between modalities, two assumptions rarely satisfied in real-world, resource-constrained environments.

Self-supervised learning (SSL) has emerged as an effective way to reduce dependence on manual annotation and improve generalisation when data are scarce. Contrastive, masked and generative strategies learn invariant multimodal representations from unlabelled data [15, 16], with contrastive objectives proving particularly effective [17, 18]. In audio-visual learning, the natural co-occurrence of sound and image provides a free supervisory signal [19, 20], and temporal synchronisation itself can serve as a training target [21]. In biological monitoring, self-supervised and few-shot methods reach robust performance with few labelled examples [22]. The application of multimodal SSL to vector surveillance nevertheless remains limited, particularly when data are heterogeneous, noisy and partially asynchronous.

Several challenges persist: temporal alignment of modalities, variable data quality and bias induced by unreliable sensors, all critical in low-resource settings [23, 24]. Dynamic fusion strategies have been proposed for low-quality multimodal data [23], and field integration has been identified as a condition for operational vector surveillance [24]. Lightweight, CPU-efficient architectures and participatory platforms favour large-scale deployment [12, 24]. Uncertainty estimation in deep learning [25] and evidential fusion that weighs each view by its estimated trustworthiness [26] offer a principled answer to unreliable sensors. Yet explicit modelling of the uncertainty attached to each modality remains largely unexplored for vector surveillance, and few approaches reconcile multimodal robustness, self-supervision, uncertainty management and real-world constraints.This study proposes an asynchronous, self-supervised and uncertainty-aware audio-visual fusion framework for vector surveillance in low-resource environments. Its objective is to learn robust multimodal representations from unlabelled data while accounting for signal asynchrony and variable sensor reliability.

The main contributions are:-

1. a self-supervised multimodal learning strategy that is robust to temporal asynchrony between audio and visual signals;
2. an adaptive fusion mechanism guided by the estimated uncertainty of each modality;
3. an experimental evaluation showing improved performance and robustness under the constraints of low-resource environments.

Materials and Methods:-

Study design:-

The framework is designed for settings characterised by low-cost sensors, noisy data and frequent lack of strict synchronisation between modalities. Unlike supervised approaches, it requires no annotation during the main learning phase and exploits only the intrinsic structure of multimodal data.

The methodology comprises five steps:-

1. acquisition and preprocessing of audio and visual data;
2. self-supervised learning of latent representations with deep encoders;
3. construction of multimodal pairs that tolerates temporal asynchrony;
4. uncertainty estimation and adaptive fusion of modalities;
5. evaluation of the learned representations on downstream tasks.

Data acquisition and preprocessing:-

The data consist of acoustic and visual signals representative of vector surveillance in low-resource environments. The acoustic modality comprises recordings of mosquito wingbeats; the visual modality comprises RGB images of mosquitoes captured under varying lighting conditions. The data come from public datasets selected to reflect realistic field acquisition scenarios.

Raw audio signals were transformed into Mel spectrograms, a standard input representation for convolutional audio classifiers [27], to capture the time-frequency characteristics of flight activity. Images were resized and normalised to reduce computational cost and to match lightweight architectures. Standard preprocessing operations were applied to mitigate variations in sensor quality and acquisition conditions. No assumption of perfect synchronisation was imposed: both streams were segmented into sliding time windows, producing loosely aligned audio-visual pairs representative of real-world misalignment.

Encoder architectures:-

Two independent encoders project each modality into a latent space: an audio encoder f_a based on a deep residual convolutional network (ResNet-18) [28] applied to Mel spectrograms, and a lightweight visual encoder f_v (MobileNetV2) [29] applied to RGB images. Two-layer multilayer perceptron (MLP) projection heads g_a and g_v map these representations into a common space where they can be aligned, a design shown to improve the quality of contrastive representations [18]. Table 1 summarises the architectures considered.

Table 1: Comparison of candidate architectures for the encoders and projection heads

Architecture	Params (≈)	FLOPs	Strengths	Limitations	Suited to low resources
ResNet-18 [28]	11 M	Moderate	Stable and robust for SSL	Less efficient than very deep models	Yes
EfficientNet-B0 [30]	5 M	Low	Good balance between accuracy and complexity	Sensitive to noisy data	Partially
Audio Transformer [31]	> 20 M	High	High representational capacity	Requires massive data	No
MobileNetV2 [29]	3.4 M	Low	Lightweight, optimised for edge devices	Less expressive than ViT	Yes
Vision Transformer (ViT) [32]	> 22 M	High	Strong global modelling	Very costly	No
MLP (2 layers, ReLU)	0.5–1 M	Low	Improves separation in the contrastive space; stabilises SSL	Slight increase in parameters	Yes

FLOPs: floating-point operations; SSL: self-supervised learning; ViT: Vision Transformer.
Given constraints on computational complexity, memory and the availability of annotated data, ResNet-18 and MobileNetV2 offer the best balance between representational robustness, stability under self-supervised training and suitability for deployment in low-resource environments.

Self-supervised contrastive objective:-
Learning follows a contrastive paradigm: the model brings together audio and visual representations of the same event and separates those of different events, without manual annotation. For an audio segment a_i and a visual segment v_j, the projected representations are $h_a = g_a(f_a(a_i))$ and $h_v = g_v(f_v(v_j))$, and their similarity is the cosine similarity $s(\cdot,\cdot)$.

Because the streams are not synchronised, every visual segment whose timestamp falls within a temporal window W of the audio segment is treated as a positive match:-

$$P_i = \{ j : | t_i^a - t_j^v | \leq W \} \quad (1)$$

The contrastive loss follows the InfoNCE formulation [17], extended to the set of admissible positives within a mini-batch B as in multi-positive contrastive learning [33], with temperature τ:-

$$\mathcal{L}_i = -\frac{1}{|P_i|} \sum_{p \in P_i} log \frac{exp(s(h_a^i, h_v^p)/\tau)}{\sum_{k \in B} exp(s(h_a^i, h_v^k)/\tau)} \quad (2)$$

This construction lets the model learn robust multimodal alignments despite temporal shifts, data loss or differences in sampling rate, which improves generalisation where precise synchronisation between sensors cannot be guaranteed.

Uncertainty estimation and adaptive fusion:-
An uncertainty module is attached to each branch, following the principle of estimating input-dependent (aleatoric) uncertainty [25]. In addition to the latent representation, it estimates an uncertainty level reflecting the reliability of the modality, which depends on factors such as acoustic noise, occlusion or low light: $\sigma_a = U_a(z_a)$ and $\sigma_v = U_v(z_v)$.

The fused representation weights each modality by a normalised confidence, in the spirit of trust-based multi-view fusion [26]:-

$$\alpha = \frac{e^{-\sigma_a}}{e^{-\sigma_a} + e^{-\sigma_v}}, \quad \beta = \frac{e^{-\sigma_v}}{e^{-\sigma_a} + e^{-\sigma_v}} \quad (3)$$

$$z_f = \alpha z_a + \beta z_v, \quad \alpha + \beta = 1 \quad (4)$$

When a modality is judged unreliable, its contribution is reduced in favour of the more informative one, yielding a joint representation adapted to degraded acquisition conditions. Fig. 1 shows the resulting architecture and Algorithm 1 the training procedure.

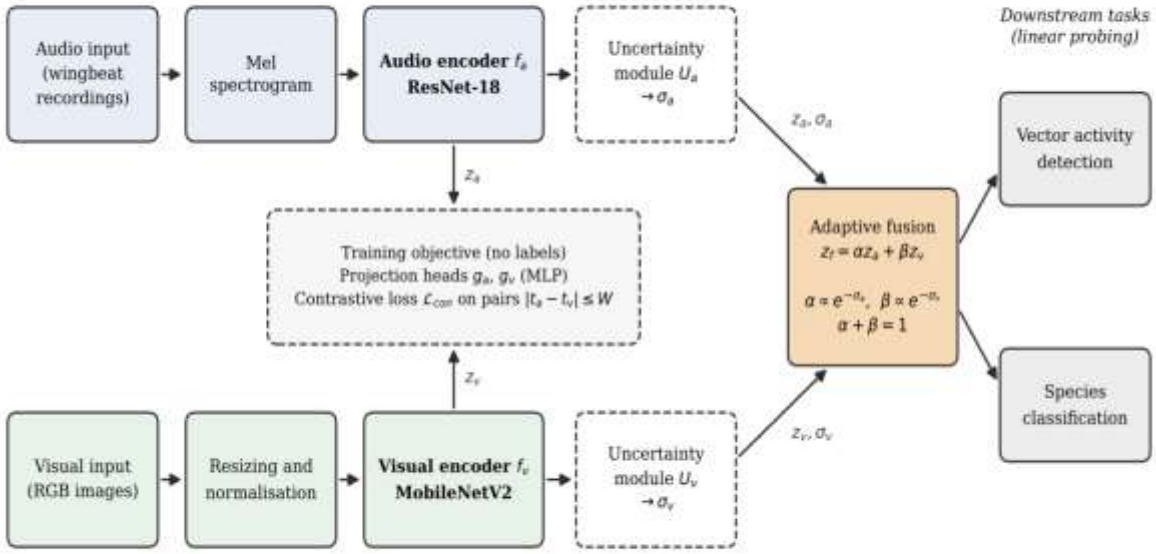


Fig. 1: Architecture of the proposed framework. Audio (Mel spectrograms) and visual (RGB images) inputs are encoded by ResNet-18 and MobileNetV2. During training, projection heads and a contrastive loss align representations of events recorded within a window W. At inference, uncertainty modules estimate σ_a and σ_v, and an adaptive fusion module weights each modality before downstream detection and classification.

Algorithm 1: Asynchronous self-supervised training with uncertainty-aware fusion

Input: unlabelled audio-visual dataset D = $\{(a_i, v_i, t_i^a, t_i^v)\}$; temporal window W; encoders f_a, f_v; projection heads g_a, g_v; uncertainty modules U_a, U_v
Output: fused multimodal representation z_f
Initialise the parameters of f_a, f_v, g_a, g_v, U_a, U_v
for each training epoch **do**
  Sample a mini-batch B from D
  for each sample (a_i, v_i) in B **do**
    Convert a_i to a Mel spectrogram
    $z_a \leftarrow f_a(a_i)$; $z_v \leftarrow f_v(v_i)$
    $h_a \leftarrow g_a(z_a)$; $h_v \leftarrow g_v(z_v)$
    Define positive pairs if $|t_i^a - t_j^v| \le W$ (Eq. 1)
    $\sigma_a \leftarrow U_a(z_a)$; $\sigma_v \leftarrow U_v(z_v)$
    $\alpha \leftarrow \exp(-\sigma_a)$; $\beta \leftarrow \exp(-\sigma_v)$
    $\alpha \leftarrow \alpha / (\alpha + \beta)$; $\beta \leftarrow 1 - \alpha$
    $z_f \leftarrow \alpha\, z_a + \beta\, z_v$
  end for
  Compute the contrastive loss L_{con} over B (Eq. 2) and update all parameters with Adam
end for
return z_f

Training procedure:-
The ResNet-18 and MobileNetV2 encoders were optimised jointly with the InfoNCE contrastive loss, using mini-batches containing loosely aligned positive pairs and negative samples drawn from distinct events. Optimisation used the Adam algorithm [34] with an initial learning rate of 1×10^{-4} and a cosine decay schedule [35] over 100 epochs. Convergence was monitored through the contrastive loss and through intermodal consistency, measured as the mean cosine similarity of positive and negative pairs.

Evaluation protocol:-
The learned representations were evaluated on downstream tasks, namely vector activity detection and species classification, by linear probing: a light linear classifier was trained on a limited annotated fraction of the dataset, the annotations being used solely for evaluation. The proposed approach was compared with an audio-only model (ResNet-18), a visual-only model (MobileNetV2) and a supervised multimodal model, in terms of accuracy, F1-score and robustness to noise.

Robustness was then assessed on data acquired in low-resource environments characterised by ambient acoustic noise, lighting variations and misalignment between sensors, without imposing any strict alignment. Three degradations were applied: acoustic noise reducing the signal-to-noise ratio (SNR) to 5 dB, low-light conditions affecting the visual modality, and random time shifts of up to ±500 ms between the audio and visual streams. Performance was measured in terms of accuracy, stability under noise and relative degradation under desynchronisation. Computational cost was measured as the mean inference time per multimodal sample on a standard CPU (four cores, 8 GB RAM).

Ethics statement:-
This study did not involve human participants or vertebrate animals. It used recordings and images of mosquitoes; no ethical approval was required.

Results:-

Training convergence:-
The contrastive loss decreased steadily throughout training, with no sign of representational collapse, indicating effective learning of the shared latent space. At the end of the 100 epochs, the mean cosine similarity reached 0.82 for positive pairs, against 0.21 for negative pairs, reflecting a clear separation in the contrastive space.

Classification performance:-
Table 2 reports linear-probing performance. The proposed approach outperformed both unimodal models and the supervised multimodal baseline, with a gain of 3.1 accuracy points over the latter.

Table 2: Classification performance by linear probing

Method	Accuracy (%)	F1-score	Noise robustness
Audio only (ResNet-18)	84.3	0.83	Moderate
Visual only (MobileNetV2)	86.1	0.85	Moderate
Supervised multimodal	89.7	0.89	Good
Proposed (SSL + uncertainty-aware fusion)	**92.8**	**0.92**	**Very good**

SSL: self-supervised learning.

Robustness under degraded conditions:-
Table 3 summarises behaviour under the three degradations. With acoustic noise at an SNR of 5 dB, the audio-only model lost 12% of its performance, whereas the proposed model lost 4%. Under low light affecting the visual modality, the uncertainty-driven weighting limited the impact of visual degradation and the accuracy remained above 88%. With random time shifts of up to ±500 ms, approaches requiring strict alignment lost up to 10%, while the extended matching and adaptive fusion of the proposed approach limited the loss to less than 3%.

Table 3: Robustness of the proposed approach under degraded acquisition conditions

Condition	Reference approach	Effect on reference approach	Effect on proposed approach
Acoustic noise, SNR = 5 dB	Audio only	−12% performance	−4% performance
Low light (visual degradation)	Visual modality	Degraded visual input	Accuracy maintained above 88%
Random time shift up to ±500 ms	Strict-alignment approaches	Up to −10% performance	Less than 3% loss

SNR: signal-to-noise ratio.

Computational efficiency:-

Inference on a standard CPU (four cores, 8 GB RAM) took 42 ms per multimodal sample on average, which is compatible with near-real-time operation on low-cost embedded devices.

Discussion:-

This study shows that combining multimodal self-supervised learning with explicit handling of temporal asynchrony and modality uncertainty improves the robustness of vector surveillance in low-resource environments. The proposed approach outperformed unimodal and supervised multimodal methods, and its advantage was most marked under degraded conditions: noise, low light and misalignment between sensors.Two mechanisms appear to drive these results. First, uncertainty-weighted fusion stabilises performance when one modality is degraded: at an SNR of 5 dB, the proposed model lost a third as much performance as the audio-only model (4% against 12%). This is consistent with work showing that dynamic fusion improves the handling of low-quality multimodal data [23], and it extends to the audio-visual case the noise robustness pursued by acoustic models such as MosquitoSong+ [7]. Second, treating every segment within a temporal window as a positive match makes the representation tolerant to the shifts produced by low-cost sensors, whereas strict-alignment approaches degrade sharply.

That a model trained without labels exceeded the supervised multimodal baseline (92.8% against 89.7%) may seem counter-intuitive. A plausible explanation is that the contrastive task, by forcing the model to associate the sound and image of the same insect, favours features shared by both modalities and discards noise specific to each, acting as a regulariser. Self-supervision can also exploit unlabelled recordings that a supervised model cannot use. These observations agree with surveys reporting better generalisation of self-supervised multimodal representations when labels are scarce [15, 16] and with results obtained in biological monitoring [22].From an operational standpoint, the 42 ms CPU inference time shows that robustness does not require specialised hardware. Together with the reduced need for expert annotation, this supports deployment in field platforms such as Ento-App [24] and in surveillance programmes where entomological expertise is limited.

This study has limitations. Although the extended matching strategy tolerates asynchrony, scenarios involving prolonged loss of a modality or progressive temporal drift may require more advanced alignment mechanisms. Uncertainty is modelled implicitly; a formal probabilistic approach, assessed with calibration measures [36], would improve the reliability and interpretability of predictions. The evaluation by linear probing still relies on an annotated fraction of the data, so annotation is reduced rather than eliminated. Finally, the experiments were conducted on public datasets and on degradations applied under controlled conditions; validation on data collected directly in operational settings, at larger scale and on field-recorded corpora such as HumBugDB [9], remains necessary.Future work will integrate additional modalities, such as environmental and geospatial variables, explore continual learning [37] to adapt the system to seasonal changes and the emergence of new vector species, and optimise the model for very low-power embedded devices to support sustainable, large-scale deployment.

Conclusion:-

This paper proposed an asynchronous, self-supervised audio-visual fusion framework with explicit uncertainty estimation for vector surveillance in low-resource environments. By combining multimodal contrastive learning, tolerance to temporal misalignment and fusion weighted by modality reliability, it learns discriminative joint representations with little reliance on manual annotation. It outperformed unimodal and supervised multimodal approaches, particularly under noise, low light and asynchrony between sensors, while remaining fast enough for low-cost hardware. By reconciling robustness, generalisation and computational efficiency, the framework

contributes to the development of intelligent and operational vector surveillance systems in constrained real-world settings.

Acknowledgments:-

The authors thank the UFR of Science and Technology of Iba Der Thiam University of Thiès for its institutional support of this doctoral research.

References:-

1. World Health Organization. World malaria report 2024: addressing inequity in the global malaria response. Geneva: World Health Organization; 2024. ISBN 978-92-4-010444-0
2. Goodwin A, Padmanabhan S, Hira S, Glancey M, Slinowsky M, Immidisetti R, et al. Mosquito species identification using convolutional neural networks with a multitiered ensemble model for novel species detection. Sci Rep. 2021;11:13656. doi:10.1038/s41598-021-92891-9
3. Trujillano F, et al. Using image segmentation models to analyse high-resolution earth observation data: new tools to monitor disease risks in changing environments. Int J Health Geogr. 2024;23(1):13. doi:10.1186/s12942-024-00371-w
4. Uelmen JA, et al. Global mosquito observations dashboard (GMOD): creating a user-friendly web interface fueled by citizen science to monitor invasive and vector mosquitoes. Int J Health Geogr. 2023;22(1):28. doi:10.1186/s12942-023-00350-7
5. Nolte K, Sauer FG, Baumbach J, Kollmannsberger P, Lins C, Lühken R. Robust mosquito species identification from diverse body and wing images using deep learning. Research Square [Preprint]. 2024. doi:10.21203/rs.3.rs-4491538/v1
6. Park J, Kim DI, Choi B, Kang W, Kwon HW. Classification and morphological analysis of vector mosquitoes using deep convolutional neural networks. Sci Rep. 2020;10:1012. doi:10.1038/s41598-020-57875-1
7. Supratak A, et al. MosquitoSong+: A noise-robust deep learning model for mosquito classification from wingbeat sounds. PLOS ONE. 2024;19(10):e0310121. doi:10.1371/journal.pone.0310121
8. Mukundarajan H, Hol FJH, Castillo EA, Newby C, Prakash M. Using mobile phones as acoustic sensors for high-throughput mosquito surveillance. eLife. 2017;6:e27854. doi:10.7554/eLife.27854
9. Kiskin I, Sinka M, Cobb AD, Rafique W, Wang L, Zilli D, et al. HumBugDB: A large-scale acoustic mosquito dataset. In: Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1 (NeurIPS Datasets and Benchmarks 2021); 2021. arXiv:2110.07607
10. Baltrušaitis T, Ahuja C, Morency LP. Multimodal machine learning: A survey and taxonomy. IEEE Trans Pattern Anal Mach Intell. 2019;41(2):423–443. doi:10.1109/TPAMI.2018.2798607
11. Nguyen CVT, Nguyen NHT, Le DT, Ha QT. Self-MI: Efficient multimodal fusion via self-supervised multi-task learning with auxiliary mutual information maximization. arXiv:2311.03785 [Preprint]. 2023. doi:10.48550/arXiv.2311.03785
12. Restrepo D, Wu C, Cajas SA, Nakayama LF, Celi LA, López DM. Multimodal deep learning for low-resource settings: A vector embedding alignment approach for healthcare applications. arXiv:2406.02601 [Preprint]. 2024. doi:10.48550/arXiv.2406.02601
13. Astruc G, Gonthier N, Mallet C, Landrieu L. OmniSat: Self-supervised modality fusion for Earth observation. In: Leonardis A, Ricci E, Roth S, Russakovsky O, Sattler T, Varol G, editors. Computer Vision – ECCV 2024. Cham: Springer Nature Switzerland; 2025. p. 409–427. doi:10.1007/978-3-031-73390-1_24
14. Irvin J, et al. USat: A unified self-supervised encoder for multi-sensor satellite imagery. arXiv:2312.02199 [Preprint]. 2023. doi:10.48550/arXiv.2312.02199
15. Tu Y. A review of the applications of self-supervised learning in multimodal models. Adv Eng Technol Res. 2025;14(1):1491. doi:10.56028/aetr.14.1.1491.2025
16. Zong Y, Mac Aodha O, Hospedales T. Self-supervised multimodal learning: A survey. arXiv:2304.01008 [Preprint]. 2024. doi:10.48550/arXiv.2304.01008
17. van den Oord A, Li Y, Vinyals O. Representation learning with contrastive predictive coding. arXiv:1807.03748 [Preprint]. 2018. doi:10.48550/arXiv.1807.03748
18. Chen T, Kornblith S, Norouzi M, Hinton G. A simple framework for contrastive learning of visual representations. In: Proceedings of the 37th International Conference on Machine Learning (ICML). PMLR 119; 2020. p. 1597–1607
19. Arandjelović R, Zisserman A. Look, listen and learn. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV); 2017. p. 609–617. doi:10.1109/ICCV.2017.73

20. Owens A, Efros AA. Audio-visual scene analysis with self-supervised multisensory features. In: Computer Vision – ECCV 2018. Lecture Notes in Computer Science, vol. 11210. Cham: Springer; 2018. p. 639–658. doi:10.1007/978-3-030-01231-1_39
21. Korbar B, Tran D, Torresani L. Cooperative learning of audio and video models from self-supervised synchronization. In: Advances in Neural Information Processing Systems 31 (NeurIPS); 2018. p. 7763–7774
22. Willi A, Baumann P, Erb S, Gröger F, Zeder Y, Lionetti S. Self-supervised and few-shot learning for robust bioaerosol monitoring. arXiv:2406.09984 [Preprint]. 2024. doi:10.48550/arXiv.2406.09984
23. Zhang Q, et al. Provable dynamic fusion for low-quality multimodal data. arXiv:2306.02050 [Preprint]. 2023. doi:10.48550/arXiv.2306.02050
24. Samb SMK, Okangondo Loshima J, Sambe TB, Diop I, Sarr MD. Ento-App: An AI-powered platform for real-time vector surveillance and epidemic forecasting. In: 2025 6th International Conference on Data Intelligence and Cognitive Informatics (ICDICI). IEEE; 2025. p. 1918–1923. doi:10.1109/ICDICI66477.2025.11134822
25. Kendall A, Gal Y. What uncertainties do we need in Bayesian deep learning for computer vision? In: Advances in Neural Information Processing Systems 30 (NeurIPS); 2017. p. 5574–5584
26. Han Z, Zhang C, Fu H, Zhou JT. Trusted multi-view classification with dynamic evidential fusion. IEEE Trans Pattern Anal Mach Intell. 2023;45(2):2551–2566. doi:10.1109/TPAMI.2022.3171983
27. Hershey S, Chaudhuri S, Ellis DPW, Gemmeke JF, Jansen A, Moore RC, et al. CNN architectures for large-scale audio classification. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2017. p. 131–135. doi:10.1109/ICASSP.2017.7952132
28. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016. p. 770–778. doi:10.1109/CVPR.2016.90
29. Sandler M, Howard A, Zhu M, Zhmoginov A, Chen LC. MobileNetV2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2018. p. 4510–4520. doi:10.1109/CVPR.2018.00474
30. Tan M, Le QV. EfficientNet: Rethinking model scaling for convolutional neural networks. In: Proceedings of the 36th International Conference on Machine Learning (ICML). PMLR 97; 2019. p. 6105–6114
31. Gong Y, Chung YA, Glass J. AST: Audio Spectrogram Transformer. In: Proceedings of Interspeech 2021; 2021. p. 571–575. doi:10.21437/Interspeech.2021-698
32. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (ICLR); 2021
33. Khosla P, Teterwak P, Wang C, Sarna A, Tian Y, Isola P, et al. Supervised contrastive learning. In: Advances in Neural Information Processing Systems 33 (NeurIPS); 2020. p. 18661–18673
34. Kingma DP, Ba J. Adam: A method for stochastic optimization. In: International Conference on Learning Representations (ICLR); 2015. arXiv:1412.6980
35. Loshchilov I, Hutter F. SGDR: Stochastic gradient descent with warm restarts. In: International Conference on Learning Representations (ICLR); 2017. arXiv:1608.03983
36. Guo C, Pleiss G, Sun Y, Weinberger KQ. On calibration of modern neural networks. In: Proceedings of the 34th International Conference on Machine Learning (ICML). PMLR 70; 2017. p. 1321–1330
37. Parisi GI, Kemker R, Part JL, Kanan C, Wermter S. Continual lifelong learning with neural networks: A review. Neural Netw. 2019;113:54–71. doi:10.1016/j.neunet.2019.01.012