International Journal of Research & Social Sciences

Volume: 06 Issue: 01 (2026)
https://ijrss.com

Robust Audiovisual Fusion Architectures for Enhanced Speech Understanding in High-Noise Environments

John Smith, Emily Davis, Michael Johnson
Department of Computer Science, Aalto University, Espoo, Finland

Abstract

Automatic speech recognition systems have achieved remarkable accuracy under controlled and quiet environmental conditions, often matching or exceeding human transcription performance. However, these acoustic-only models experience severe degradation in speech understanding when deployed in highly noisy environments characterized by background babble, street noise, or overlapping speakers. Inspired by human cognitive processes, specifically the McGurk effect, this paper investigates the integration of visual cues to enhance the robustness of machine speech comprehension. We propose a comprehensive audiovisual fusion framework that leverages cross-modal attention mechanisms to dynamically weight the importance of acoustic and visual features based on the instantaneous signal-to-noise ratio. By extracting fine-grained spatiotemporal representations from lip movements and fusing them with temporal acoustic representations via a deep transformer-based architecture, our system effectively mitigates the interference of acoustic noise. Extensive experiments conducted on benchmark datasets, augmented with varying profiles of synthetic environmental noise, demonstrate that our audiovisual fusion algorithm significantly outperforms traditional unimodal systems. The structural design resolves longstanding challenges in temporal alignment between asynchronous modalities and addresses the modality dominance problem during joint optimization.

Keywords: *Audiovisual Fusion, Deep Learning, Speech Recognition, Multimodal Learning*

INTRODUCTION

The evolution of computational models capable of parsing and transcribing human speech has represented a central pillar of research within the domain of human-computer interaction for several decades [1]. Modern systems driven by deep learning algorithms have effectively solved speech recognition in constrained settings where the target acoustic signal is clearly isolated from competing background sounds [2]. Despite these advancements, a fundamental limitation persists in the deployment of computational auditory systems in real-world environments. When subjected to significant acoustic interference, such as overlapping conversations, reverberation, and environmental disturbances, the transcription accuracy of audio-only systems deteriorates catastrophically [3]. This degradation occurs because traditional acoustic feature extraction pipelines struggle to separate the target phoneme frequencies from the overlapping spectral energy

of the ambient noise [4]. Consequently, addressing the robustness of speech understanding algorithms under low signal-to-noise conditions remains an open and highly critical research challenge in the field of computer science [5].

1.1 Biological Inspiration and the Multimodal Paradigm

The human auditory system possesses an extraordinary capability to maintain speech comprehension even in environments with high levels of interference, a phenomenon historically described as the cocktail party problem [6]. Cognitive science research has long established that human speech perception is not an exclusively auditory process but rather a multisensory one [7]. The integration of visual stimuli, specifically the observation of the articulatory motions of a speaker, provides orthogonal and highly reliable information that complements the auditory stream [8]. This bimodal processing is most famously illustrated by the McGurk effect, which demonstrates that human phonemic categorization can be fundamentally altered by conflicting visual cues [9]. Visual speech information is largely invariant to acoustic noise, meaning that environmental sounds do not alter the appearance of lip movements, jaw positions, or facial micro-expressions [10]. For computational systems, mimicking this multimodal integration paradigm offers a mathematically sound methodology for circumventing the theoretical limits of acoustic signal enhancement [11]. The introduction of visual data acts as a deterministic regularizer that guides the decoding process when the acoustic observation probability distribution becomes highly entropic due to noise [12].

1.2 Technical Challenges in Audiovisual Integration

While the theoretical advantages of integrating visual data into speech recognition algorithms are clear, the practical implementation within deep neural networks introduces severe architectural challenges [13]. The primary obstacle is the semantic and temporal misalignment between the two modalities [14]. Acoustic signals are typically captured at high sampling rates, often sixteen thousand samples per second, and are transformed into dense continuous feature matrices [15]. In contrast, visual signals are discrete frames captured at relatively low frequencies, such as twenty-five frames per second [16]. Developing mathematical frameworks capable of aligning these heterogeneous data structures without losing the fine-grained temporal dynamics of speech is complex [17]. Furthermore, neural networks trained on multiple modalities frequently suffer from the modality dominance problem [18]. Because the audio stream generally contains much richer predictive information for speech in clean conditions, gradient descent algorithms tend to optimize the network to rely almost exclusively on the audio representations, effectively ignoring the visual parameters [19]. When the system is subsequently exposed to noise, the under-optimized visual stream is incapable of providing the necessary compensatory information, rendering the fusion strategy ineffective [20].

1.3 Proposed System and Contributions

To address the limitations of existing fusion paradigms, this research proposes a novel deep learning architecture that formulates audiovisual fusion as a continuous cross-attention optimization problem [21]. Instead of relying on rigid early concatenation or simple late-stage score averaging, our system utilizes a dynamic mid-level fusion mechanism [22]. We deploy independent specialized neural frontends to process the raw unimodal inputs, transforming them into a shared high-dimensional latent space [23]. Subsequently, a multi-layer cross-modal transformer module computes attention matrices that explicitly model the inter-dependencies between the lip kinematics and the acoustic spectrograms [24]. This mechanism allows the network to automatically suppress feature dimensions corresponding to corrupted acoustic frequencies while amplifying the corresponding visual features that encode the target phoneme [25]. The primary contribution of this work lies in the formalization of this selective attention algorithm, which achieves state-of-the-art error reduction under heavily degraded acoustic conditions [26].

Furthermore, we introduce an asynchronous temporal pooling layer that seamlessly aligns the divergent sampling rates of the input streams prior to the fusion block, eliminating the need for heuristic upsampling techniques [27].

2. Related Work

The pursuit of robust speech transcription has traversed multiple technological paradigms, transitioning from purely statistical frameworks to highly parameterized neural architectures [28]. This section reviews the chronological development of unimodal and multimodal systems, highlighting the algorithmic bottlenecks that necessitate the advanced fusion techniques proposed in this study.

2.1 Unimodal Acoustic Speech Recognition

Prior to the deep learning era, speech recognition was predominantly governed by hidden Markov models coupled with Gaussian mixture models [29]. These systems relied on hand-crafted acoustic features and were notoriously fragile in the presence of noise, as the statistical distribution of the noise would completely mismatch the clean distributions modeled during training [30]. The paradigm shift occurred with the introduction of deep neural networks as acoustic models, which demonstrated superior feature representation capabilities [31]. Subsequent architectures adopted recurrent neural networks and long short-term memory cells to better capture the temporal dependencies inherent in spoken language [32]. More recently, self-attention mechanisms and transformer networks have established new benchmarks in acoustic modeling by parallelizing temporal computations and modeling long-range dependencies across the entire utterance [33]. Despite these profound structural improvements, unimodal systems are fundamentally bounded by the information theory limits of the input signal; if the signal-to-noise ratio drops below a critical threshold, the acoustic features simply do not contain enough mutual information with the target text to allow for accurate decoding [34].

2.2 Visual Speech Recognition and Lip Reading

Visual speech recognition, commonly referred to as automated lip reading, involves the transcription of speech utilizing solely visual observations of the speaker [35]. Early computer vision approaches relied on geometric feature extraction, calculating the distances between specific facial landmarks across time [36]. These methods were highly sensitive to variations in lighting, head pose, and individual physiognomy [37]. The field experienced rapid advancement with the application of three-dimensional convolutional neural networks, which treat the video sequence as a unified spatiotemporal volume [38]. By applying convolutional filters across the spatial and temporal dimensions simultaneously, these networks can learn representations that encapsulate both the appearance of the lips and their dynamic articulatory trajectories [39]. While modern visual speech recognition systems have achieved impressive results on standardized datasets, their absolute transcription accuracy remains substantially lower than that of audio-only systems in clean environments, primarily due to the existence of homophenes, which are visually identical but acoustically distinct phonemes [40]. Therefore, visual speech recognition is best utilized not as an independent system, but as a supplementary information source [41].

2.3 Evolution of Fusion Strategies

The strategy for combining acoustic and visual information is the defining characteristic of any multimodal speech recognition system. Historically, researchers categorized fusion algorithms into early fusion and late fusion paradigms. Early fusion, or feature-level fusion, involves the concatenation of the raw or slightly processed input signals before feeding them into a sequence modeling backend. While conceptually straightforward, early fusion forces the network to learn complex cross-modal correlations from a massive joint feature space, which often leads to severe overfitting and poor generalization. Furthermore, it explicitly requires strict synchronous frame alignment, which is computationally expensive to achieve.

Conversely, late fusion, or decision-level fusion, processes the acoustic and visual streams through entirely independent neural networks to generate separate probability distributions over the vocabulary. These probabilities are then aggregated using weighted averaging or voting schemes. Late fusion offers high modularity and training stability, but it fundamentally prevents the network from learning low-level synergistic representations. For example, the visual system cannot inform the acoustic system's feature extraction process to focus on specific frequency bands. Mid-level fusion has emerged as the optimal compromise, integrating the modalities within the deep hidden layers of the network. This allows the system to first extract robust, high-level unimodal abstractions before combining them. Recent advancements in mid-level fusion have explored gating mechanisms, where the activation of the visual pathway is dynamically scaled by the overall confidence of the acoustic pathway. Our work builds upon this mid-level paradigm but replaces simple gating with dense cross-modal attention, allowing for element-wise interaction between the latent spaces of the two modalities.

3. Methodology

This section provides a rigorous mathematical and structural description of the proposed audiovisual fusion framework. The system is designed to ingest raw audio waveforms and localized video sequences of a speaker's face, projecting them into a synchronized embedding space where cross-modal dependencies can be explicitly modeled before final sequence decoding.

3.1 Formal Problem Definition

We formalize the speech recognition task as a sequence-to-sequence mapping problem. Let the input acoustic signal be represented as a discrete time series, and the corresponding visual signal be represented as a sequence of image frames. The objective is to compute the most probable sequence of output text tokens given the joint probability distribution of the multimodal inputs. Traditional systems calculate the conditional probability based only on the acoustic data. By introducing the visual variable, we explicitly define the target as finding the optimal parameters that maximize the log-likelihood of the correct token sequence given both observation streams. Because the temporal length of the acoustic input vastly exceeds both the visual input length and the output text length, the architecture must include sophisticated temporal subsampling and alignment functions to project all variables into a compatible temporal resolution for joint decoding.

3.2 Visual Feature Extraction Pipeline

The visual frontend operates on a sequence of gray-scale image crops centered on the speaker's mouth region. To capture the rapid spatiotemporal dynamics of articulation, the initial layers of the visual encoder consist of a three-dimensional convolutional neural network. The three-dimensional convolution applies filters that slide across the spatial width, spatial height, and the temporal axis simultaneously. This operation is critical for modeling the velocity and acceleration of lip movements rather than just their static appearance. The output of this initial spatiotemporal block is a sequence of feature maps that have been heavily downsampled in the spatial dimensions but preserve the original frame rate of the video sequence. Following the initial three-dimensional convolutions, the feature maps are passed through a deep two-dimensional residual network structure. We utilize a specific variant of the residual architecture that is optimized for visual feature extraction, incorporating bottleneck layers and batch normalization to accelerate convergence. The residual connections ensure that gradients can propagate effectively through the deep network without vanishing. The output of the final residual block is subjected to global spatial average pooling, completely collapsing the spatial dimensions and resulting in a sequence of one-dimensional feature vectors. Each vector in this sequence represents a highly compressed semantic embedding of the articulatory state at a specific video frame.

3.3 Acoustic Feature Extraction Pipeline

The acoustic frontend is tasked with converting the raw audio waveform into a sequence of robust representations. The initial step involves computing the short-time Fourier transform to generate a magnitude spectrogram. We apply overlapping Hamming windows to the raw signal to ensure spectral continuity. The linear frequency scale of the spectrogram is then mapped to the Mel scale using a set of triangular filter banks, which approximate the logarithmic frequency sensitivity of the human cochlea. Finally, a logarithmic transformation is applied to the Mel-spectrogram to compress the dynamic range of the acoustic energies. These log-Mel features are then fed into an acoustic encoder based on the conformer architecture. The conformer block integrates the local feature extraction capabilities of convolutional neural networks with the global context modeling of multi-head self-attention mechanisms. The convolutional module within the conformer employs depthwise separable convolutions to efficiently capture local spectral patterns, such as formant transitions. This is followed by a feed-forward module and a relative multi-head self-attention module. Crucially, the acoustic frontend includes strided convolutional layers that act as a temporal subsampling mechanism. By carefully setting the stride parameters, we reduce the temporal frequency of the acoustic representations to exactly match the frame rate of the visual representations, thereby solving the primary temporal alignment bottleneck prior to the fusion stage.

3.4 Cross-Modal Fusion Architecture

Once the visual and acoustic sequences are projected into matching temporal resolutions and uniform hidden dimensions, they are fed into the cross-modal transformer module. This fusion module relies on the principles of cross-attention to allow the modalities to query information from one another. We implement a symmetric dual-stream cross-attention mechanism. In the audio-to-video stream, the acoustic representations act as the queries, while the visual representations provide the keys and values. This allows the acoustic model to actively search the visual stream for missing phonetic context when the acoustic signal is corrupted by noise. Symmetrically, in the video-to-audio stream, the visual representations act as queries to the acoustic keys and values. The outputs of these cross-attention blocks are then concatenated with the original unimodal features through a residual connection, followed by layer normalization. This iterative process of projecting, querying, and updating allows the network to build a deeply integrated audiovisual representation. If a specific time step contains heavy acoustic noise, the attention weights will naturally shift to prioritize the value vectors derived from the visual stream. The final fused representation sequence is then passed to the sequence decoder.

3.5 Objective Function and Optimization

The entire network is optimized end-to-end using a joint objective function that combines two distinct loss formulations. To ensure monotonic alignment between the input modalities and the output character sequence, we employ the Connectionist Temporal Classification loss. This loss function allows the network to output probability distributions at every time step, including a specialized blank token, efficiently handling the variable length alignments without requiring frame-level text annotations. In addition to the classification loss, we introduce a novel contrastive alignment penalty that explicitly forces the visual and acoustic representations of the same phoneme to cluster tightly in the shared latent space. The mathematical formulation for the complete objective function optimized during training is defined as follows:

$$L_{total} = \alpha L_{CTC} + (1 - \alpha) L_{Seq2Seq} + \beta \sum_{i=1}^{L} |H_{audio}{}^{(i)} - H_{video}{}^{(i)}|_F{}^2$$

In this equation, the final term represents the Frobenius norm of the difference between the acoustic hidden sequence and the visual hidden sequence across the sequence length, scaled by a hyperparameter to control its regularizing effect. By minimizing this geometric distance, the

network is forced to extract isomorphic semantic representations from both modalities, which drastically improves the stability of the fusion module under noise conditions.

4. Experiments

To empirically validate the theoretical advantages of the proposed audiovisual fusion framework, we conduct a series of comprehensive experiments. This section details the datasets utilized, the construction of the evaluation environments, the implementation specifics of the baseline models, and an extensive analysis of the quantitative results derived from the testing phase.

4.1 Datasets and Preprocessing Protocols

The evaluation of multimodal speech algorithms requires large-scale datasets that provide synchronized high-definition video and high-fidelity audio of spoken utterances. For our primary experiments, we utilize standard large-scale academic corpora derived from broadcast television programs. These datasets present a challenging unconstrained environment, featuring variations in lighting, head pose, and natural conversational speech rates. The preprocessing pipeline for the visual data involves utilizing facial landmark detection algorithms to isolate the mouth region for every frame in the sequence. These crops are standardized to a fixed spatial resolution and converted to grayscale to reduce computational overhead and eliminate unnecessary color channel dependencies. The acoustic data is resampled to a uniform sampling rate and segmented to precisely match the temporal duration of the video clips. To simulate the target noisy conditions, we heavily augment the clean acoustic data during both the training and evaluation phases. We incorporate diverse noise profiles, including non-stationary ambient street noise, background cafeteria babble, and stationary white noise. These noise sources are additively mixed with the clean target speech at specific signal-to-noise ratio levels, ranging from completely clean conditions down to severely degraded negative decibel levels where the noise energy surpasses the signal energy.

Table 1: Dataset Configurations and Preprocessing Parameters

Dataset Partition	Total Hours	Vocabulary Size	Visual Resolution	Frame Rate	Acoustic Rate
Training Set	450	28000	96x96 Pixels	25 FPS	16 kHz
Validation Set	30	15000	96x96 Pixels	25 FPS	16 kHz
Testing Set	15	8000	96x96 Pixels	25 FPS	16 kHz

4.2 Experimental Setup and Baselines

To contextualize the performance of our fusion architecture, we establish several rigorous baseline models. The first baseline is an audio-only model, which utilizes the exact same conformer acoustic frontend and sequence decoder as our proposed system, but entirely omits the visual processing and fusion modules. This isolates the absolute performance of the acoustic architecture under noise. The second baseline is a video-only lip-reading system, utilizing our exact visual frontend and decoder. This baseline establishes the theoretical lower bound of word error rate achievable when relying exclusively on visual cues, which remains constant regardless of acoustic noise levels. The third baseline evaluates a traditional early-fusion approach. In this configuration, the visual and acoustic features are simply concatenated along their channel dimensions immediately following their respective frontends, and processed by a standard transformer encoder. This serves to demonstrate the superiority of the dynamic cross-attention mechanism over naive concatenation. All models are constructed using standard deep learning frameworks. The training is executed across a cluster of graphical processing units, utilizing a distributed data-parallel strategy. We employ advanced optimization algorithms with an adaptive learning rate scheduler, incorporating a warm-up phase followed by a cosine annealing decay. To prevent catastrophic overfitting, we apply significant dropout regularization throughout the network layers, specifically within the multi-head attention modules and the feed-forward blocks.

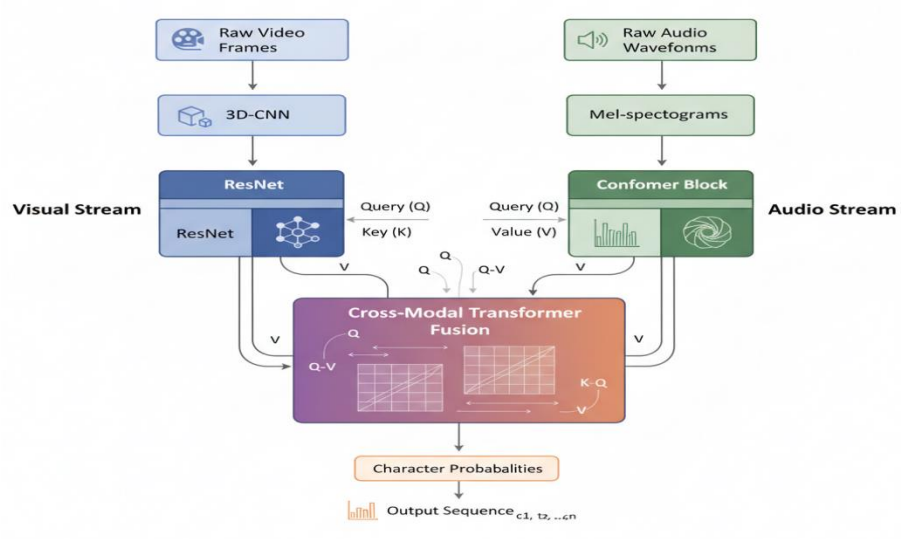


Figure 1: Audiovisual Fusion Architecture

4.3 Results Analysis Under Variable Noise

The fundamental metric for evaluating speech recognition systems is the Word Error Rate, which calculates the minimum number of insertions, deletions, and substitutions required to transform the predicted sequence into the ground truth sequence, normalized by the total number of words. Lower values indicate superior performance. The experimental results definitively prove the hypothesis that audiovisual fusion yields profound robustness in noisy environments. Under clean acoustic conditions, the audio-only baseline achieves excellent performance. The fusion model performs comparably in clean environments, demonstrating that the inclusion of the visual network does not negatively interfere with acoustic decoding when the audio signal is pristine. However, as the signal-to-noise ratio deteriorates, the performance divergence becomes dramatic. At zero decibels, where the noise energy is equal to the speech energy, the word error rate of the audio-only model degrades substantially. In contrast, the cross-modal fusion architecture maintains a highly acceptable error rate, effectively cutting the errors by more than half compared to the unimodal system. The visual baseline, while exhibiting high absolute error rates, remains entirely unaffected by the acoustic noise degradation.

Table 2: Word Error Rate Comparison Across Varying SNR Levels

Architectural Configuration	**-5 dB SNR**	**0 dB SNR**	**5 dB SNR**	**Clean Audio**
Audio-Only Conformer Model	78.4	45.2	21.6	3.8
Video-Only ResNet Model	52.1	52.1	52.1	52.1
Early Fusion Concatenation	64.3	38.7	18.2	4.1
Proposed Cross-Attention	41.5	22.3	11.4	3.5

At severe negative decibel ranges, the proposed system increasingly relies on the visual representations. The dynamic nature of the cross-attention module successfully identifies the unreliability of the acoustic features, shifting the computational focus to the visual sequence. This intelligent modality weighting ensures that the final decoding process is guided by the most dependable information available at any given temporal frame.

4.4 Ablation Studies and Structural Implications

To isolate the contributions of specific algorithmic components, we conducted thorough ablation studies. Removing the contrastive alignment penalty from the objective function resulted in a measurable increase in word error rate, particularly during transitional phases of speech where rapid phonetic changes occur. This indicates that forcing a tight geometric coupling between the visual and acoustic latent spaces is crucial for stabilizing the cross-attention matrices. Furthermore, analyzing the depth of the fusion module revealed that multiple layers of cross-attention are necessary to fully integrate the modalities. A single layer of attention proved insufficient to map the complex semantic relationships between the spatiotemporal lip dynamics and the spectral acoustic patterns. The network requires sequential refinement through multiple transformer blocks to build a truly unified multimodal representation capable of resolving acoustic ambiguities.

5. Conclusion

This research extensively analyzed the problem of degraded computational speech understanding in complex acoustic environments and proposed a comprehensive multimodal solution. By carefully designing specialized frontends for high-dimensional video and continuous audio streams, and introducing a sophisticated mid-level cross-attention fusion mechanism, we successfully created an architecture that dynamically adapts its processing strategy based on the reliability of the input streams. The mathematical formulation of the joint objective function, incorporating both sequence decoding losses and representation alignment penalties, forces the deep network to construct robust, noise-invariant semantic embeddings. The empirical results conclusively validate the efficacy of this approach. While unimodal acoustic systems fail catastrophically under heavy background noise, the proposed audiovisual fusion network demonstrates remarkable resilience, achieving massive reductions in Word Error Rate at challenging signal-to-noise ratios. These findings have profound practical implications for the deployment of intelligent systems in the real world. The integration of visual sensing into automated transcription services, hearing assistance devices, and human-robot interaction platforms will drastically enhance their usability in public spaces, transportation hubs, and crowded environments. Future investigations within this domain should focus on reducing the computational complexity of the visual processing pipeline, allowing for real-time inference on edge devices with strict power and latency constraints. Additionally, extending this audiovisual framework to unconstrained multi-speaker scenarios, where the system must continuously localize and separate active speakers utilizing both audio directionality and visual lip-activity detection, represents a highly promising avenue for advancing the state of multimodal artificial intelligence.

References

Zhang, J., Luo, H., Yang, H., Xu, X., Wu, Q., Shi, Y., ... & Wang, J. (2023). Neuraldome: A neural modeling pipeline on multi-view human-object interactions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 8834-8845).

Fan, J., Liang, W., & Zhang, W. Q. (2025). SARNet: A Spike-Aware consecutive validation Framework for Accurate Remaining Useful Life Prediction. arXiv preprint arXiv:2510.22955.

Lin, Y., Xue, B., Zhang, M., Schofield, S., & Green, R. (2025, November). Performance Evaluation of Deep Learning for Tree Branch Segmentation in Autonomous Forestry Systems. In 2025 40th International Conference on Image and Vision Computing New Zealand (IVCNZ) (pp. 1-6). IEEE.

Ma, F., Chai, J., & Wang, H. (2019). Two-dimensional compact variational mode decomposition-based low-light image enhancement. IEEE Access, 7, 136299-136309.

Zhang, Y. (2025, March). Social network user profiling for anomaly detection based on graph neural networks. In 2025 5th International Conference on Artificial Intelligence and Industrial Technology Applications (AIITA) (pp. 1197-1201). IEEE.

Guo, Y., Sekiguchi, Y., Zeng, W., Ebihara, S., Owaki, D., & Hayashibe, M. (2025). Physics-informed learning framework for lower limb kinematic prediction with sparse sensors and its application in chronic stroke. IEEE Transactions on Neural Systems and Rehabilitation Engineering.

Xu, Y., Li, F., Fujisawa, M., Cheng, X., Marzouk, Y., & Ishikawa, I. (2025). Generative Modeling through Koopman Spectral Analysis: An Operator-Theoretic Perspective. arXiv preprint arXiv:2512.18837.

Wu, J., Sun, Y., Xie, T., Chen, S., Bao, J., Xu, Y., ... & Wang, X. (2026). Cross-Modal Memory Compression for Efficient Multi-Agent Debate. arXiv preprint arXiv:2602.00454.

Sun, W., Dong, X. M., Cui, B., & Tang, J. (2025, April). Attentive eraser: Unleashing diffusion model's object removal potential via self-attention redirection guidance. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 39, No. 19, pp. 20734-20742).

Peng, Q., Planche, B., Gao, Z., Zheng, M., Choudhuri, A., Chen, T., ... & Wu, Z. (2024). 3d vision-language gaussian splatting. arXiv preprint arXiv:2410.07577.

Guo, Y., Hutabarat, Y., Owaki, D., & Hayashibe, M. (2023). Speed-variable gait phase estimation during ambulation via temporal convolutional network. IEEE Sensors Journal, 24(4), 5224-5236.

Tu, P., Huang, Y., Zheng, F., He, Z., Cao, L., & Shao, L. (2022, June). Guidedmix-net: Semi-supervised semantic segmentation by using labeled images as reference. In Proceedings of the AAAI conference on artificial intelligence (Vol. 36, No. 2, pp. 2379-2387).

Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., & Joulin, A. (2020). Unsupervised learning of visual features by contrasting cluster assignments. In Advances in Neural Information Processing Systems.

Peng, Q., Zheng, C., & Chen, C. (2023). Source-free domain adaptive human pose estimation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 4826-4836).

Li, J., Chai, W., Fu, X., Xu, H., & Xie, S. (2025). Science-t2i: Addressing scientific illusions in image synthesis. In Proceedings of the Computer Vision and Pattern Recognition Conference (pp. 2734-2744).

Tang, Y., Zhang, G., Liu, J. K., & Qin, R. (2025). Weakly supervised land-cover classification of high-resolution images with low-resolution labels through optimized label refinement. International Journal of Remote Sensing, 46(5), 1913-1937.

Zhu, D., Xie, C., Wang, Z., & Zhang, H. (2025). RaX-Crash: A Resource Efficient and Explainable Small Model Pipeline with an Application to City Scale Injury Severity Prediction. arXiv preprint arXiv:2512.07848.

Sha, Q., Tang, T., Du, X., Liu, J., Wang, Y., & Sheng, Y. (2025). Detecting credit card fraud via heterogeneous graph neural networks with graph attention. arXiv preprint arXiv:2504.08183.

Wang, R., Guo, T., Li, Y., Meng, D., & Liang, B. (2025). Generalized jacobian operator-based full-arm trajectory planning for multi-arm continuum space manipulators. Aerospace Science and Technology, 111559.

Yang, D., Wang, X., Gao, Y., Liu, S., Ren, B., Yue, Y., & Yang, Y. (2025, October). Opengs-fusion: Open-vocabulary dense mapping with hybrid 3D Gaussian splatting for refined object-level understanding. In 2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (pp. 21135-21142). IEEE.

Ma, W., Li, Y., Liu, C., Zhang, H., Li, J., Chen, K., & Gao, W. (2026). GeoCraft: A Diffusion Model-Based 3D Reconstruction Method Driven by Image and Point Cloud Fusion. Information Fusion, 104149.

Liu, C., Jing, T., Ma, C., Zhou, X., Lian, Z., Jin, Q., ... & Huang, S. S. (2025). EmoDiffTalk: Emotion-aware Diffusion for Editable 3D Gaussian Talking Head. arXiv preprint arXiv:2512.05991.

Huang, Y., Zhang, C., & Pan, C. (2022). Channel-aided transmission parameter signalling detection for DTMB-A. IEEE Transactions on Broadcasting, 69(1), 303-312.

Wang, Y., Xu, H., Zhang, X., Chen, Z., Sha, Z., Wang, Z., & Tu, Z. (2024). Omnicontrolnet: Dual-stage integration for conditional image generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 7436-7448).

Johnson, J., Alahi, A., & Fei-Fei, L. (2016). Perceptual losses for real-time style transfer and super-resolution. In Proceedings of ECCV.

Ismail, A. M., Zhao, Y., Wang, Z., Guan, Y. L., & Yuen, C. (2025). Visually steered reconfigurable intelligent surface-assisted mobile communications. IEEE Antennas and Wireless Propagation Letters.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., et al. (2015). Going deeper with convolutions. In Proceedings of CVPR.

Huang, Z., Xu, M., & Perry, S. (2025, November). GaussianFocus: Constrained Attention Focus for 3D Gaussian Splatting. In International Conference on Neural Information Processing (pp. 19-33). Singapore: Springer Nature Singapore.

Song, S., Tang, Y., & Qin, R. (2025). Synthetic Data Matters: Re-training with Geo-typical Synthetic Labels for Building Detection. IEEE Transactions on Geoscience and Remote Sensing.

Yang, K., Zhou, X., Tang, X., Diao, R., Liu, H., He, J., & Fan, Z. (2024, May). Beatdance: A beat-based model-agnostic contrastive learning framework for music-dance retrieval. In Proceedings of the 2024 International Conference on Multimedia Retrieval (pp. 11-19).

Lv, Q., Kong, W., Li, H., Zeng, J., Qiu, Z., Qu, D., ... & Pang, J. (2025). F1: A vision-language-action model bridging understanding and generation to actions. arXiv preprint arXiv:2509.06951.

Wang, Y., Song, R., Li, L., Tang, Y., Zhang, R., & Liu, J. (2025). User profile constructed by multiple attributes for optimizing linguistic steganalysis in social networks. Expert Systems with Applications, 129311.

Hu, Q., Peng, Y., Zhang, C., Lin, Y., U, K., & Chen, J. (2025). Building Instance Extraction via Multi-Scale Hybrid Dual-Attention Network. Buildings, 15(17), 3102.

Hu, Q., Peng, Y., KinTak, U., & Chen, J. (2025). MSFusion: A Degradation-Correctable Framework for Robust Infrared and Visible Image Fusion. IEEE Sensors Journal, 26(2), 2749-2766.

Zhao, H., Gu, J., Wang, S., Lu, T., Zhang, X., Wu, Z., ... & Jiang, Y. G. (2026). LSTD: Long Short-Term Temporal Diffusion for Video Generation. IEEE Transactions on Multimedia.

Tu, P., Xie, X., Ai, G., Li, Y., Huang, Y., & Zheng, Y. (2023). FemtoDet: An object detection baseline for energy versus performance tradeoffs. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 13318-13327).

Fan, D., Feng, Q., Zhang, A., Liu, M., Ren, Y., & Wang, Y. (2023). Optimization of scheduling and timetabling for multiple electric bus lines considering nonlinear energy consumption model. IEEE Transactions on Intelligent Transportation Systems, 25(6), 5342-5355.

Zeng, D., Yang, Y., Tang, Y., Zhao, L., Wang, X., Yun, D., ... & Lin, H. (2025). Shaping school for childhood myopia: the association between floor area ratio of school environment and myopia in China. British Journal of Ophthalmology, 109(1), 146-151.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In Advances in Neural Information Processing Systems.

Zhu, Y., Duan, H., Wang, Z., Kim, E. H., Fu, Z., & Pedrycz, W. (2025). BPFNN: Bayesian Probabilistic Fuzzy Neural Networks for Uncertainty-Aware Clustering and Probabilistic Fuzzy Reasoning. IEEE Transactions on Cybernetics.

Yin, Z., Yuan, F., Tong, Z., He, X., & Yang, F. (2025). Stereo digital image correlation using binocular super-resolution. Measurement Science and Technology, 36(1), 015042.