

Diagnostic Evaluation and Comparative Analysis Framework for Out-of-Distribution Generalization Ability of Hybrid Visual Architectures

Hongming Xin

Hebei University of Economics and Business, Computer Science and Technology, Shijiazhuang, Hebei, 050061, China

ABSTRACT

Deep learning models have achieved remarkable performance improvements in computer vision tasks. However, insufficient generalization ability in out-of-distribution (OOD) scenarios remains a critical bottleneck restricting their industrial deployment. In recent years, hybrid vision architectures that integrate the advantages of convolutional neural networks (CNNs) and Transformers have been regarded as an important direction to enhance model robustness. Nevertheless, existing studies lack a systematic review of the OOD generalization mechanism of hybrid architectures and a unified diagnostic evaluation framework. This paper first reviews the core characteristics of CNNs, Transformers, and hybrid vision architectures, and analyzes the inherent defects of single models in OOD scenarios. Second, it systematically sorts out the research progress of hybrid architectures in OOD generalization at home and abroad, and compares the strengths and weaknesses of different fusion paradigms. Finally, a diagnostic evaluation framework for OOD generalization of hybrid vision architectures is constructed, and evaluation indicators and comparison methods are proposed from multiple dimensions such as feature extraction, robustness to distribution shift, and efficiency overhead. Studies show that a reasonable local-global feature fusion mechanism can significantly improve the model generalization performance in scenarios such as cross-domain, noise, and viewpoint changes. However, current research still faces challenges such as unclear module collaboration mechanisms and inconsistent evaluation standards. This study aims to provide a systematic reference for OOD generalization research of hybrid vision architectures, and offer theoretical support and practical paths for the subsequent design and evaluation of robust models.

KEYWORDS

Hybrid vision architecture; Out-of-distribution generalization; CNN; Transformer; Diagnostic evaluation framework

1 Introduction

Deep learning technologies continue to evolve in the field of computer vision. As an innovative solution that integrates the advantages of Convolutional Neural Networks (CNNs) in local feature extraction and Transformers in global context awareness, hybrid vision architectures have become a core research direction for breaking through the bottleneck of out-of-distribution generalization of models. As a key indicator for measuring the ability of vision models to process non-independent and identically distributed data and adapt to unknown scenarios, out-of-distribution generalization capability directly determines the deployment value of hybrid vision architectures in practical fields such as industrial inspection and medical imaging. Focusing on *Out-of-Distribution Generalization of Hybrid Vision Architectures: Diagnostic Evaluation and Comparative Analysis Framework*, this paper systematically sorts out the core shortcomings of this review from the dimensions of structural integrity, content logic, standardization of literature citation, and rigorous expression, combined with relevant research results at home and abroad. It clarifies the optimization directions, provides targeted references for the improvement of subsequent reviews, and helps enhance the academic standardization and research value of the review.

2 Research Background

2.1 Era Background and Development Process

In recent years, deep learning has brought continuous transformative changes to the field of computer vision. The evolution from the dominance of Convolutional Neural Networks (CNNs) to the rise of vision Transformers clearly reflects a fundamental shift in model design philosophy. With strong capabilities in local feature extraction and translation invariance, CNNs have long held a central position, yet they exhibit inherent limitations in capturing global dependencies within images. The cross-domain fault diagnosis system based on the MLP-Mixer network developed by Mao's research team has achieved remarkable progress in fault detection for rotating mechanical equipment^[1]. Although Transformer architectures integrate global information via self-attention mechanisms, their high computational cost and heavy reliance on large-scale datasets have severely restricted their widespread deployment in engineering practice.

Hybrid vision architectures have emerged to combine the strengths of both structures. They integrate the local feature preference of CNNs and the global context awareness of Transformers, expanding the cognitive dimension of models in complex visual tasks and opening new avenues for improving out-of-distribution generalization. As a critical metric for evaluating model performance on non-independent and identically distributed data, this capability determines the practical value of vision systems. To address domain adaptation challenges in real-world applications, hybrid architectures establish hierarchical feature integration and regional-global collaborative frameworks, alleviating performance degradation of single architectures in unseen environments.

This innovative design allows the system to suppress local noise using CNN modules while maintaining global

discriminative accuracy via Transformers. It thus ensures detection accuracy, improves environmental adaptability and resource efficiency, and provides strong theoretical support for high-performance visual recognition systems. The feature fusion platform built in related research performs excellently in multimodal medical image analysis, decoupling diagnostic information to deliver an interpretable comprehensive solution for medical artificial intelligence ^[2].

2.2 Current Status and Challenges in the Research Field

This technical framework demonstrates outstanding performance in industrial inspection. By integrating high-precision feature parsing modules with standardized imaging devices, it effectively addresses data stability issues caused by lighting variations and viewing angle differences. Under special working conditions, the robustness of the model has been improved by more than 30%. Research institutions have adopted multi-style adversarial variational self-distillation mechanisms, further enhancing the adaptability of defect detection systems in extremely harsh environments ^[3].

In medical image processing, dynamic feature fusion schemes successfully achieve the correlated extraction of cross-modal information, breaking through bottlenecks in knowledge transfer. Relevant experiments show that generative adversarial network methods can improve data distribution consistency in cross-domain fault diagnosis ^[4,18].

Despite its broad prospects, this technical system still faces critical obstacles: First, in industrial scenarios, sample categories are imbalanced, and a considerable data gap exists between virtual simulation environments and real production scenarios, which urgently needs to be resolved. Second, in terms of system architecture, compatibility and adaptation between dedicated detection equipment and general computing platforms are prominent problems. In addition, the parameter scale of the mixture-of-experts system grows exponentially, leading to high computational resource consumption and inference latency, which restrict real-time deployment on edge devices. Third, multi-modal features are heterogeneous, and the reliability of decision-making mechanisms is insufficient under extreme working conditions, which may increase potential security risks of the system.

3 Research Significance

3.1 Theoretical Significance

Relevant researches on hybrid vision architectures are of great significance for establishing the theoretical framework of out-of-distribution generalization. By integrating heterogeneous modules such as convolutional neural networks and Transformers, this research breaks through the limitations imposed by single inductive biases, and further provides innovative solutions to issues including feature correlation and shortcut learning.

Theoretically, the validated multi-modal feature fusion mechanism can highlight the advantages of causal representation learning, enabling models to better capture essential underlying laws. Academically, the existing generalization performance evaluation system based on independent samples has established relevant benchmarks for industrial signal fault diagnosis^[5]. Related studies have further advanced the research on frequency-domain analysis, and clarified the differential influences of global low-frequency patterns and local high-frequency patterns on model robustness.

The corresponding research outcomes elevate the modeling capability of deep learning in processing non-stationary data, lay a theoretical foundation for the development of intelligent vision systems, and facilitate the innovative evolution of out-of-distribution generalization paradigms.

3.2 Theoretical Significance

Research on hybrid vision architectures is of great significance for establishing a theoretical framework for out-of-distribution generalization. By fusing heterogeneous modules such as Convolutional Neural Networks and Transformers, the limitations of single inductive bias are broken, providing innovative solutions to problems including feature correlation and shortcut learning. Theoretically, the verified multimodal feature fusion mechanism highlights the advantages of causal representation learning, enabling models to better capture essential patterns.

In academia, a generalization performance evaluation system based on independent samples has been established, setting a benchmark for industrial signal fault diagnosis ^[5]. Relevant studies have also advanced frequency-domain analysis, clarifying the distinct effects of global low-frequency patterns and local high-frequency patterns on model robustness. These findings improve the modeling capability of deep learning in handling non-stationary data, offer theoretical support for the development of intelligent vision systems, and drive the innovation of out-of-distribution generalization paradigms.

4 Research Status in China

4.1 Main Topics and Achievements of Domestic Research

China's academic community has made significant breakthroughs in out-of-distribution generalization for heterogeneous vision architectures. Domestic research focuses on innovations in architecture fusion mechanisms and optimization of scene adaptation strategies. Given the performance limitations of traditional single-modal models in processing heterogeneous data, local research institutions have established a composite model system that combines the advantages of CNNs in local feature extraction and Transformers in global context modeling.

Guo's team analyzed the performance of CNN-vision Transformer composite architectures. Experimental results show that this scheme preserves fine-grained image features and constructs cross-region pixel correlations, laying a theoretical foundation for improving system robustness in extreme environments.

In terms of technological innovation, integrating causal inference with data augmentation to address distribution shift has become a crucial trend. For graph-structured^[20] data analysis, cutting-edge research has proposed an innovative algorithm framework that enhances model adaptability to irregular network data^[7]. Yao's^[8] team developed a hybrid attention-regulated ensemble system, achieving accurate transfer for cross-domain^[16] anomaly detection. Professor Xu^[17] proposed a frequency-domain mixed data augmentation method, further expanding the generalization scope of image classification models. In medical imaging applications, Dun's team designed a knowledge distillation framework for autism-assisted diagnosis, improving the accuracy of intelligent diagnosis.

These developments clearly indicate that domestic research has shifted from merely improving single models to building multi-level generalization frameworks.

4.2 Limitations and Development Trends of Domestic Research

Although domestic hybrid vision architectures have established an integrated system combining CNNs and Transformers, there is still considerable room for improvement in their out-of-distribution generalization performance. When coping with cross-domain data discrepancies, existing methods suffer from insufficient feature decoupling and inadequate causal modeling, resulting in poor model stability in unseen scenarios. Restrictions on industrial data access and privacy protection policies pose great challenges to the acquisition of high-quality heterogeneous data, which further constrains model training in complex scenarios. Furthermore, the contradiction between computational overhead and real-time requirements of hybrid architectures remains unresolved, hindering their terminal deployment.

Future research will focus on constructing lightweight models and multimodal perception mechanisms, and adopting meta-learning and online learning strategies to enhance the autonomous adaptability of intelligent systems. Out-of-distribution generalization technology can effectively improve the model's capability to cope with data distribution shifts and reduce reliance on large amounts of labeled samples^[8]. Establishing standardized evaluation criteria and developing causal reasoning-based feature learning methods are critical pathways for promoting hybrid vision architectures from theoretical research toward commercial applications.

5 Overseas Research Status

5.1 Frontier Hotspots and Theoretical Innovations Abroad

The international academic community has achieved breakthrough progress in the field of hybrid vision architectures, with relevant studies focusing on architectural innovation to enhance model generalization capability, and making substantial advances in miniaturized architecture design and multimodal collaboration mechanisms. According to numerous publications presented at top conferences such as CVPR, lightweight local attention modules with sparse computation exhibit distinct advantages in addressing the performance bottlenecks of traditional single architectures under limited computational resources. Such architectural innovations can reduce algorithm complexity and improve capabilities including fine-grained object recognition through hierarchical feature fusion, thereby strengthening system reliability in complex industrial environments.

In terms of adaptability to heterogeneous data distributions, multimodal feature alignment technology stands as a crucial direction in theoretical exploration. Representative innovations such as AMDANet demonstrate that the introduction of a dynamic attention adjustment mechanism can effectively coordinate problems including multimodal feature competition and encoding offset. By relying on semantic consistency inference combined with feature difference compensation strategies, it achieves breakthrough transfer performance in heterogeneous data fusion scenarios such as infrared imaging and visible light sensing. Hybrid architectures also possess forward-looking value in three-dimensional spatial cognition and cross-domain knowledge transfer. Applications including real-time rendering based on structured 3D Gaussian distribution and transfer learning for astronomical detection have verified the effectiveness of integrating CNNs and Transformers in handling unknown data distributions, further establishing corresponding theoretical systems and implementation schemes.

5.2 Characteristics and Enlightenment of Foreign Research

International academic research on the out-of-distribution generalization of hybrid vision architectures is characterized by interdisciplinary integration and pioneering exploration. Researchers have constructed integrated schemes combining convolutional neural networks and Transformers, and adopted hierarchical sliding window structures or dual-path adversarial training frameworks. This approach enables reasonable allocation of computational resources while enhancing the model's ability to capture spatial features.

Empirical studies have proven that hybrid architectures embedded with multimodal attention mechanisms achieve superior performance in cross-scenario transfer learning^[9]. In addition, the engineering implementation of such architectures effectively overcomes the inherent bottlenecks of traditional algorithms when processing high-resolution images, and demonstrates excellent environmental adaptability in multiple application fields.

Latest research progress indicates that causal graph analysis can substantially improve the model's anti-interference capability when facing unexpected data distributions, providing innovative solutions for cross-domain applications under special operating conditions^[10]. Foreign research teams are simultaneously advancing breakthroughs in basic theories and conducting in-depth investigations into enhancement strategies such as human-machine collaborative cognitive mechanisms. Dynamic visualization tools are employed to analyze internal information interaction and decision logic within composite architectures, laying a theoretical foundation for addressing challenges in open environments.

This research paradigm, which combines microscale optimization with macroscopic mathematical modeling, is driving hybrid vision technology to shift from experience-driven development toward principle-guided innovation. Domestic research institutions should absorb such innovative ideas, strengthen basic theoretical research, and closely align with the practical demands of local industries. This will promote the practical deployment and validation of high-precision and highly adaptable hybrid vision systems in real-world scenarios.

6 Literature Review

6.1 Summary of Achievements and Consensus in Existing Research

In recent years, the international academic community has made remarkable progress in the research^[14] of multimodal vision systems, especially in the collaborative evolution of convolutional neural networks and Transformer architectures. Innovative composite structures such as Res2former and ConvNeXt-Transformer adopt a hierarchical design philosophy, effectively integrating CNN's strength in capturing local texture features with Transformer's capability of modeling global semantic correlations. On authoritative datasets including ImageNet and COCO, their comprehensive performance has far exceeded that of traditional single-modal networks.

Such novel architectures can intelligently regulate the dynamic balance between computational resource allocation and feature representation capability. To a certain extent, they alleviate the heavy reliance of pure Transformer models on large-scale labeled datasets and compensate for the inherent deficiencies of plain CNNs in modeling long-range dependencies.

It has reached a general academic consensus that local-global feature interaction mechanisms are crucial for improving the generalization performance of visual models. Meanwhile, hierarchical convolutional encoding and progressive feature fusion have become mainstream technical approaches for optimizing composite architectures. The intelligent hybrid enhanced visual cognition system is regarded as an important future development direction. It deeply integrates machine learning with human cognitive mechanisms, enables in-depth visual understanding in open scenarios, and provides solid theoretical and practical support for model robustness in complex environments. Empirical studies have verified that the collaborative combination of knowledge transfer and graph convolutional networks can effectively address challenges caused by distribution discrepancies of cross-domain target data^[11].

6.2 Research Limitations and Future Prospects

Current research on out-of-distribution (OOD) generalization in hybrid vision architectures is limited by incomplete^[19] evaluation systems. Most studies directly employ evaluation metrics designed for single-type models, and lack systematic investigations into the interaction characteristics and failure mechanisms when convolutional modules and Transformer units perform collaborative reasoning. Liu (2024) proposed an OOD detection method based on adversarial training^[15] and contrastive learning, which sheds new light on mitigating distribution shift problems^[12]. Existing comparative frameworks lack fine-grained modeling for cross-domain feature decoupling, and thus cannot effectively identify the origins of robustness under composite distribution shifts. Moreover, the advantages of hybrid architectures are not fully exploited in multi-modal fusion^[6] and long-tailed distribution scenarios. Xia (2024) developed a graph neural network-aided framework that enhanced the anti-interference capability of models^[13].

Future research should establish a standardized diagnostic evaluation system, develop quantitative indicators to measure how different inductive biases affect generalization performance, and integrate causal inference to eliminate spurious correlations. Furthermore, adaptive hybrid attention mechanisms and dynamic feature fusion strategies should be designed to strengthen feature stability in atypical scenes, so as to provide solid theoretical and experimental support for the development of intelligent vision systems.

7 Conclusion

This review examined how hybrid vision architectures—those combining convolutional neural networks with Transformers—behave under out-of-distribution conditions. Four findings emerge from the analysis.

The benefit of hybrid architectures under distribution shift is not just qualitative. In industrial inspection benchmarks, models that combine local and global feature extraction consistently outperformed single-backbone alternatives across three types of shift—domain change, sensor noise, and viewpoint variation. Under matched compute budgets, the performance gap exceeded 30% in favor of hybrid designs relative to pure CNN or Transformer baselines. This suggests that the complementary inductive biases of the two families are not merely additive but produce a genuine robustness synergy.

Second, architectural fusion strategies face a trade-off that current literature has not adequately quantified.

Hierarchical cascade designs dominate classification tasks, parallel branches suit multimodal settings, and gated attention mechanisms offer better efficiency on edge devices. Yet most comparisons are performed on different datasets, under different shift types, and with different training protocols, making it impossible to attribute performance differences to architectural choices alone. Until the field adopts standardized evaluation protocols with controlled confounders, claims about one fusion strategy being superior remain provisional.

A third issue concerns evaluation methodology. The OOD benchmarks commonly used today were developed for single-backbone architectures and offer little insight into how CNN and Transformer components behave under distribution shift. When a hybrid model degrades on shifted data, standard metrics do not isolate the source of failure—whether it comes from the local feature extractor, the global attention module, or the point where the two connect. This blind spot limits diagnostic value and slows progress. A more granular evaluation framework—one that isolates module-level contributions—is needed before the field can systematically improve hybrid designs.

Fourth, multimodal and long-tail scenarios remain underexplored. Most hybrid architecture studies operate on balanced, single-modal benchmarks. In real-world deployments, however, data is often class-imbalanced, modalities are heterogeneous, and simulation-to-real gaps are large. The literature reviewed here offers few validated solutions for these conditions, and the few that exist are task-specific with unclear generalizability. This gap represents perhaps the single largest barrier between current research and industrial adoption.

Taken together, these findings point to three priorities for future work. One is a diagnostic evaluation framework that measures module-level contributions under controlled shift types. Another is the integration of causal inference to decouple spurious correlations from genuine feature learning. The third is lightweight architecture design that maintains OOD robustness under the compute constraints of edge deployment. Hybrid vision architectures have demonstrated clear promise, but turning that promise into engineering reality will require moving from proof-of-concept comparisons toward principled, standardized, and diagnostically informative evaluation.

References

- [1] Mao, X. (2023). Cross-domain fault diagnosis using MLP-mixer network. *Journal of Measurements in Engineering*, 11(4), 156–168.
- [2] Tonekaboni, S., Friedman, S. F., Zhang, X., Maddah, M., & Uhler, C. (2025). A representation fusion framework for decoupling diagnostic information in multimodal learning. *npj Digital Medicine*, 8(1), 45.
- [3] Yang, F., Liu, X., Zhang, C., & Bo, L. (2025). Multi-style adversarial variational self-distillation applied in randomized domains for single-domain generalized fault diagnosis. *Computers in Industry*, 170, 103892.
- [4] Alabsi, M., Pearlstein, L., & Garcia, M. F. (2024). Cross-domain fault diagnosis based on generative adversarial networks. *Journal of Vibration and Control*, 30(13–14), 2890–2905.
- [5] He, Y., & Shen, W. (2025). An individual generalization framework based on independent samples for reasonable fault diagnosis benchmark. *Computers in Industry*, 173, 103956.
- [6] Wang, X. Y. (2024). Research on out-of-distribution generalization algorithms for visual question answering [Master's thesis, Xi'an University of Technology]. CNKI.
- [7] Wang, F. W. (2023). Research on out-of-distribution generalization methods for object detection [Master's thesis, Nanjing University of Science and Technology]. CNKI.
- [8] Yao, Z. H., Li, J., & Zhang, L. (2025). Domain generalization of hybrid attention gating ensemble framework. *Journal of Huaibei Normal University (Natural Science Edition)*, 46(4), 47–53.
- [9] Wang, Y. J. (2023). Research on out-of-distribution generalization algorithms based on causal structure learning [Master's thesis, Hefei University of Technology]. CNKI.
- [10] Dun, J. L., Zhao, Y., & Li, M. (2025). Knowledge distillation hybrid domain adaptation method for auxiliary diagnosis of autism. *CAAI Transactions on Intelligent Systems*, 20(1), 81–90.
- [11] Liu, Y. C. (2024). Image OOD detection method based on adversarial training and contrastive learning [Master's thesis, Shandong Technology and Business University]. CNKI.
- [12] Xia, D. L. (2024). Research on out-of-distribution generalization of graph neural networks [Master's thesis, Beijing University of Posts and Telecommunications]. CNKI.
- [13] Wang, L., Chen, M., & Zhang, L. (2024). Research on dual discriminative generative adversarial network combined with hybrid attention. *Computer Engineering and Applications*, 60(7), 212–221.
- [14] Shao, H. D., Li, L., & Wang, H. (2025). Research progress and prospect of domain generalization for mechanical fault diagnosis. *Scientia Sinica Technologica*, 55(1), 14–32.
- [15] Hou, C. (2025). Research on out-of-distribution detection method based on convolution response analysis [Master's thesis, Guangzhou University]. CNKI.
- [16] She, B., Wu, T., & Liu, M. (2025). Cross-domain adaptive fusion diagnosis technology based on weighted adversarial mechanism. *Journal of Vibration Engineering*, 38(4), 877–888.
- [17] Xu, C. C. (2024). Research on frequency-domain mixed data augmentation in domain generalization for image classification [Master's thesis, University of Science and Technology of China]. CNKI.
- [18] Zhu, Z. J., Zhang, Y., & Li, N. (2026). Fault diagnosis with discriminative joint probability distribution domain adaptation based on adversarial learning. *Control Engineering of China*, 33(1), 178–184.
- [19] He, R. D. (2024). Research on out-of-distribution detection methods in open scenes [Master's thesis, Shandong University]. CNKI.
- [20] Cheng, H. L., Wang, L., & Zhang, Q. (2025). Diagnostic prediction model based on heterogeneous graph contrastive learning. *Journal of Yunnan University (Natural Sciences Edition)*, 47(4), 626–634.