Proceedings of CONF-MLA 2025 Symposium: Applied Artificial Intelligence Research
DOI: 10.54254/2755-2721/2025.BJ25969

Local-Global Synergy: The Architectural Evolution and Paradigm Fusion of CNN-Transformer Hybrid Models in the Field of Computer Vision

Kairui Liu

Manchester Metropolitan Joint Institution, Hubei University, Wuhan, China
202231123001082@stu.hubu.edu.cn

Abstract. Convolutional Neural Networks (CNNs) excel at local feature extraction but lack global scope, while Vision Transformers (ViT) capture global context but are computationally expensive and lack crucial inductive biases. To resolve this trade-off, CNN-Transformer hybrid models have emerged to synergize these strengths, becoming a dominant architectural paradigm in computer vision. However, a comprehensive analysis of their evolutionary trajectory and the profound "spillover" of their core "local-global synergy" philosophy is lacking. This paper provides a systematic review of this evolution, charting its development through four key stages: (1) early modular splicing and replacement, (2) native synergistic design for unified architectures, (3) ideological fusion influencing pure CNN and Transformer paradigms, and (4) the current rise of "Poly-Hybrids" integrating emerging operators like State-Space Models (SSMs). We analyze the critical challenges confronting these advanced models, including prohibitive resource barriers, interpretability black boxes, and the fine-grained alignment gap in vision-language tasks. We conclude that the field is at an inflection point, where the pursuit of "stronger" models must yield to the necessity of "more reliable" ones. Future progress will hinge not just on performance, but on achieving breakthroughs in sustainability, trustworthiness, and alignment, positioning these architectures as the perceptual bedrock for general world models.

Keywords: CNN-Transformer Hybrid Models, Local-Global Synergy, Architectural Evolution, Paradigm Fusion, Computer Vision

1. Introduction

Nowadays, modern computer vision mostly adopts two main methods. Convolutional Neural Networks (CNNs) are very good at extracting small, local details because they have built-in rules such as locality and translation invariance [1]. However, when CNNs stack many layers, it is still difficult for them to see the entire image or the far-flung connected parts [2]. The creation of the Vision Transformer (ViT) solves this problem by using the "self-attention" tool to instantly view all parts of the image and understand large images [3]. But ViT does not have the same built-in rules as CNNs, which makes it perform poorly on relatively small data sets without a large amount of pre-training. In addition, processing high-resolution images requires a lot of computer energy [3].

In order to solve this trade-off problem, the CNN-Transformer hybrid model came into being. These models attempt to combine CNN 's techniques with local details and Transformer 's capabilities

DOI: 10.54254/2755-2721/2025.BJ25969

with large graphs, achieving a good balance between accuracy, data efficiency, and computational cost [4]. Although there are many hybrid models, few people have studied how they change over time and how the main ideas affect other areas. This article wants to fill this gap. We will describe how these models have evolved, starting with simple blending, then moving to better design, and finally showing how they affect pure models and new AI systems.

Our work has mainly done two things. In theory, we provide a clear and comprehensive view of these changes, which can bring some inspiration to future design. In practice, this work provides a guide for model developers to help them discover great challenges. At the same time, it also points out the future direction for establishing a more accurate and reliable visual model.

2. The architectural evolution of CNN-Transformer hybrid models

2.1. Early exploration: concatenation and replacement as independent modules

In the beginning, early hybrid models simply mixed or swapped CNN and Transformer parts. They treated these parts as separate "black boxes." This process begins with a non-intrusive "auxiliary improvement" approach, such as architectures like non-local networks [5] inserted self-attention blocks in parallel to supplement the global acceptance field of the CNN. Later, this concept gradually evolved into a more radical "structural replacement", where models like BoTNet [6]and DETR [7], which serially replace the core convolutional layers with Transformer components or directly replace the entire downstream head. Although the design of this early stage is basically limited to the assembly of existing modules, it is crucial to verify the hybrid concept and establish the basic combination mode, which lays the foundation for further local integration.

2.2. Maturing: native synergy and unified design

In the following stage, the design of the model began to change from repairing the existing model to designing the local collaborative architecture from scratch. Researchers began to regard convolution and self-focusing as equally important "building materials" to create new architectures. This idea has spawned some influential models, including CoAtNet [4] and Next-ViT [8], which use different or interleaved convolution and attention blocks. These architectures jointly validate the key rules of hybrid design, namely, the efficiency of using CNN to process shallow local details and the power of attention to process deep global semantics. In addition, this natively designed model has a strong multi-task generalization ability, and models such as UniFormer [9] seamlessly process image and video tasks within a single framework. So far, the CNN-Transformer hybrid architecture has been established as a mature and independent architecture category, and the ability of local-global collaboration has been raised to a new height.

2.3. Ideological spillover and paradigm fusion

The success of the hybrid model is not limited to the field of hybrid architecture. Its core concept of "local-whole collaboration" has begun to produce "spillover" and reshape the broader field. This fusion of ideas clearly shows a two-way feature. On the one hand, the pure Transformer architecture, like Swin Transformer [10], absorbs the concepts of CNN's hierarchy and locality to enhance its visual modeling capabilities. Conversely, pure CNNS like ConvNeXt [11] systematically adopt the design principles of Transformer, such as reverse bottlenecks and large cores, and have achieved decent performance without any concern mechanisms. This indicates that excellent design principles can break through the boundaries of operators. Elevate "hybrid" from a physical concept to an ideological one.

DOI: 10.54254/2755-2721/2025.BJ25969

This philosophical quickly permeated into more complex fields. In the field of generative AI, models like LaVin-DiT [12] adopted a macro-level hybrid synergy, that is, they used CNN-like encoder for efficient representation while using Transformer core for generation. Rely on such a way to achieve large-scale and multi-tasking applications. In multimodal systems, this thinking reshapes the interaction between vision and language. Models like ParGo [13] and HunyuanCustom [14] do more than just mix features. They deeply connect the parts that "see" and the parts that "create". Now, this "hybrid" idea is a key way to build big AI systems.

2.4. The rise of "Poly-Hybrids"—fusing with emerging paradigms

Now, we have 'Poly-Hybrids'. These new models only use CNN and Transformer architectures. They add new ways to calculate and make wise choices. A big change is the addition of state space models (SSMs) and fast calculation steps. Models such as MambaVision [15] and ViG [16] are early examples. They combine CNNs, Transformers and SSMs to work as a team. This team approach saves a lot of time and effort when viewing very large pictures.

At the same time, these models use rules to determine what to calculate. They no longer just blindly process data. They are able to choose what to do. For example, SeqMvRL [17] uses reinforcement learning to select the best path and mixed facts for tasks with multiple views. These new skills make Poly-Hybrids the main tool for building top-level media and generating applications today. For example, LaVin-DiT [12] and HunyuanCustom [14] link how artificial intelligence "sees" and how to "create". This proves that mixing different parts is a very useful and flexible way to build modern AI.

3. Core challenges

"Poly-Hybrid" has excellent performance and efficiency and can do many different tasks. But their complex structures also bring new, intractable problems. These problems are not just mathematical problems. They also involve how much hardware resources we need, how to maintain the security of generated things, and emotional ethics.

3.1. Challenge one: sustainability and the resource barrier

At present, in order to pursue excellent performance, the advanced Poly-Hybrids training process requires huge hardware resource consumption. Taking the latest Mamba Vision as an example, running 300 epochs on ImageNet-1K requires 32 × A100 GPUs [15]. For more complex generative basic models, resource consumption will only be more. Building a unified model like LaVin-DiT that can handle more than 20 tasks requires a total of 120,000 steps of training on 64 A100-80G GPUs, relying on complex distributed training techniques such as DeepSpeed ZeRO-2 [12]. Such a huge computing demand not only brings high economic costs and increased carbon emissions, but also most small and medium-sized research institutions and enterprises do not have sufficient hardware resources, making it extremely difficult to replicate and innovate cutting-edge models, hindering the popularization and democratization of technology.

In the current era of "Poly-Hybrids" this challenge is exacerbated by its inherent complexity. First, fragmented software-hardware co-optimization arises from the mix of heterogeneous operators. The distinct computational profiles of self-attention (IO-intensive), SSMs (requiring specialized kernels), and dynamic routing (causing load imbalance) make automatic compiler fusion and optimization inefficient [18-21]. Secondly, the hidden cost of data processing has soared. The demand for large-scale, noisy, web crawler data sets such as LAION-5B [22] and WebVid-2M [23] involves complex and expensive processing workflows, which may exceed the cost of computing itself and will also become a major bottleneck for sustainable development.

DOI: 10.54254/2755-2721/2025.BJ25969

3.2. Challenge two: the interpretability black box and security risks

If a single CNN or Transformer is already a "black box" the multi-operator, multi-path nature of Poly-Hybrids creates a "deeper black box" severely hindering trust and transparency. Traditional Attribution Methods, such as Grad-CAM [24], which is designed to visualize model decisions, often fail or produce misleading results when applied to these complex architectures. It becomes nearly impossible to intuitively determine whether a final decision was dominated by a CNN's local textures, an SSM's long-range memory, a Transformer's global semantics, or their non-linear interactions. This ambiguity in the internal decision-making logic greatly undermines our trust in the model. This ambiguity is unacceptable in high-stakes applications like autonomous driving and medical image diagnosis, we must be able to trace its erroneous attribution, which module, which feature led to this misjudgment? If it cannot be explained, the same type of risk cannot be fixed or avoided.

This opacity directly gives rise to novel security risks. While adversarial attacks on single-paradigm models are well-documented [25], the intricate structure of Poly-Hybrids introduces new, poorly understood vulnerabilities. Attackers could move beyond simple input perturbations and instead target the model's internal logic. For example, a carefully crafted attack might exploit the dynamic routing mechanism in ParGo [13] to force suboptimal computational paths, or manipulate the RL-based decision process in SeqMvRL [17] to corrupt its fused representation. These attacks on "structural weaknesses" could be far more covert and destructive than traditional methods, yet the community's understanding of how to defend against them is still in its infancy.

3.3. Challenge three: fine-grained vision-language alignment

Modern Poly-Hybrids like HunyuanCustom [14] and ParGo [13] have achieved great success in empowering Large Vision-Language Models (LVLMs). However, a fundamental challenge has emerged: in complex vision-language processing workflows, while models can achieve semantic understanding at a macroscopic level, they often fail at the microscopic level where Fine-Grained Alignment is required. For example, a model might correctly identify "a cat is playing the piano" but provide a wrong answer when asked, "Which key did the cat's left paw press?" This inconsistency between macro-level scene understanding and micro-level detail localization leads to a prevalent issue of "Semantic Hallucination" where the generated text has a factual misalignment with the actual visual content. This challenge has become one of the core bottlenecks for large vision-language models [26].

This consistency challenge is particularly prominent nowadays. Its root cause lies in many aspects. First of all, after going through that rather complex processing, the output of the visual backbone is mostly a flat feature vector, which lacks a clear object-centered structure. Just as a recent survey on object-centered learning has emphasized, explicitly modeling visual entities and the hierarchical relationships between them is the key to achieving a more robust and interpretable visual understanding [27]. Secondly, the projection layer that connects vision and language may become a new source of distortion. ParGo [13]'s research shows that if it is poorly designed, it will magnify detail errors, causing the illusion of high confidence. The noise that already exists in the training data is like the noise in LAION-5B [22]. Deep models process this noise, which accumulates and magnifies, thereby fundamentally increasing the alignment difficulty.

This kind of failure in fine-grained alignment will directly limit the reliable application of poly-hybrid in safety-critical fields. In the scenario of autonomous driving, if there is a misunderstanding of complex traffic instructions, it may cause serious accidents. In the field of robotics, incorrect alignment will hinder the execution of precise operation tasks. This situation is particularly evident in cutting-edge fields like Vision-Language Navigation (VLN), as enabling agents to accurately follow navigation instructions in unfamiliar real-world environments is highly dependent on high-

DOI: 10.54254/2755-2721/2025.BJ25969

quality visual-language alignment capabilities [28]. Narrowing this consistency gap has become a core technical challenge in driving hybrid models from "powerful" to "reliable". And it can be truly deployed in high-risk scenarios.

4. Future outlook

4.1. Towards a dynamic and unified architecture

Future hybrid architectures will no longer be fixed, static combinations but rather dynamic, composable "compute engines" that achieve ultimate "on-demand synergy." Early signals of this trend have been shown in Par Go's Dynamic Projector and Seq MvRL's Reinforcement Learning View Selector [13,17]. The future model architecture will bring this idea to the extreme. An architecture may have an "expert database" containing multiple operators such as CNN, SSM and Attention. The higher-level "meta-controller" will dynamically combine and invoke these experts according to the input in real time to match the optimal local or global processing power for different regions of the image or different stages of the task. The latest research of Vision Mixture-of-Experts (V-MoE), from the pioneering V-MoE [20] to the recently optimized training stability ViT-MoE, is actively exploring how to efficiently train and deploy such dynamic routing systems. This transition from static coordination at design time to dynamic coordination at inference time will be the key to achieving final computational efficiency.

4.2. Towards green, trustworthy, and responsible AI

As the capacity of the model grows exponentially, the need for its sustainability and credibility will also become the focus of research. In the future, the evaluation criteria of SOTA model will be multi-dimensional, including not only accuracy, but also training energy consumption, interpretability, authentication robustness boundary and algorithm fairness measurement. The work represented by Green NAS [28] has begun to incorporate training energy consumption (kWh) into the neural architecture search target, providing a feasible template for "green training". The new generation of interpretable methods represented by ViT-CX [29] attempts to establish an explanatory framework that is more faithful to the model decision logic through techniques such as causal inference. This is the first step towards fine-grained robustness assessment and trust building. These efforts jointly promote the community to compare "energy-interpretability-fairness" with traditional accuracy indicators, and provide an operable tool chain and evaluation paradigm for transforming "powerful" multi-mixture into "reliable" social tools.

4.3. As the perceptual bedrock for general world models

The ultimate task of the visual model is to provide a perceptual basis for Artificial General Intelligence (AGI). The "outside world" will increasingly constrain their design. As Yang Likun advocated in the paper, the perception module must output an object-centered structured representation suitable for reasoning [30]. A recent authoritative survey on object-centered learning has also shown that explicit modeling of visual entities and their hierarchical relationships, and direct feedback of these objectified tokens to language or action planning heads, can significantly improve the accuracy and robustness of multi-step reasoning [27].This suggests that the optimization objective for hybrid backbones will shift from ImageNet classification accuracy to whether they can efficiently extract invariances from the physical world (such as causality, geometry, and physical laws) and encode them into structured knowledge that can be understood by higher-level decision-making and planning models. The design of visual architectures will increasingly shift from being "data-driven" to being "physics- and task-driven."

DOI: 10.54254/2755-2721/2025.BJ25969

5. Conclusion

This paper has systematically reviewed the evolutionary journey of CNN-Transformer hybrid models in the field of computer vision. Their core idea of "local-global synergy" has undergone a clear, deepening evolutionary path: from the early concatenation and replacement of physical modules, to the maturation of native synergistic architectures in the middle stage, and then to a profound ideological spillover into pure paradigms, ultimately evolving in the current stage into an open and powerful Poly-Hybrid that actively integrates State-Space Models (SSM), dynamic mechanisms, and multimodal demands.

Even though "Poly-Hybrids" are very successful, they now face a turning point. They must deal with high costs, the hidden dangers of the "black box," and the problem of matching small details in vision and text. These big problems show that just trying to make models bigger and faster is not enough anymore. The rewards for doing this are getting smaller. Because of this, we think the next big step for hybrid models is to change the main goal from making them "Stronger" to making them "More Reliable".

The best model in the future will make significant improvements in three aspects: saving energy, obtaining trust and correctly matching data. In order to do this, researchers need to stop just building a single network. They should focus on creating an open, less-powerful, trusted vision system. This means that we must build systems that can change and work together to run at a very fast pace. We also need to ensure that these visual models are secure enough to be the eyes of a larger artificial intelligence system. Moreover, we must create new ways to test them, to see how much power they use, if we can understand their choices, if they are fair. Using "local-global collaboration" is just a tool. The real goal is to build AI that we can actually use in a cluttered real world.

References

[1] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR).

[2] Luo, W., Li, Y., Urtasun, R., & Zemel, R. (2016). Understanding the effective receptive field in deep convolutional neural networks. In Advances in neural information processing systems (NeurIPS).

[3] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. In International Conference on Learning Representations (ICLR).

[4] Dai, Z., Liu, H., Le, Q. V., & Tan, M. (2021). CoAtNet: Marrying Convolution and Attention for All Data Sizes. arXiv preprint arXiv:2106.04803.

[5] Wang, X., Girshick, R., Gupta, A., & He, K. (2018). Non-local Neural Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 7794–7803).

[6] Srinivas, A., Lin, T. Y., Parmar, N., et al. (2021). Bottleneck Transformers for Visual Recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 16519–16529).

[7] Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). End-to-end object detection with transformers. In European conference on computer vision (ECCV) (pp. 213-229). Springer, Cham.

[8] Jiashi Li, et al. "Next-ViT: Next Generation Vision Transformer for Efficient Deployment in Realistic Industrial Scenarios." arXiv 2022.

[9] Li, K., Wang, Y., Zhang, J., He, Y., Li, H., Xue, F., & Wang, J. (2022). UniFormer: Unifying convolution and self-attention for visual recognition. arXiv preprint arXiv:2201.09450.

[10] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, 10012-10022.

[11] Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., & Xie, S. (2022). A ConvNet for the 2020s. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, 11976-11986.

[12] Wang, Z., Xia, X., Chen, R., Yu, D., Wang, C., Gong, M., & Liu, T. (2025). LaVin-DiT: Large Vision Diffusion Transformer. Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 20060–20070.

DOI: 10.54254/2755-2721/2025.BJ25969

[13] Wang, A.-L., Shan, B., Shi, W., Lin, K.-Y., Fei, X., Tang, G., Liao, L., Tang, J., Huang, C., & Zheng, W.-S. (2025, April 11). ParGo: Bridging Vision-Language with Partial and Global Views. Proceedings of the 39th Annual AAAI Conference on Artificial Intelligence.https://doi.org/10.1609/aaai.v39i7.32806
[14] Hu, T., Yu, Z., Zhou, Z., Liang, S., Zhou, Y., Lin, Q., & Lu, Q. (2025). HunyuanCustom: A Multimodal-Driven Architecture for Customized Video Generation. arXiv preprint arXiv:2505.04512v2. https://doi.org/10.48550/arXiv.2505.04512
[15] Hatamizadeh, A., & Kautz, J. (2025). MambaVision: A Hybrid Mamba-Transformer Vision Backbone. Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 25261-25270.
[16] Liao, B., Wang, X., Zhu, L., Zhang, Q., & Huang, C. (2025). ViG: Linear-complexity Visual Sequence Learning with Gated Linear Attention. Proceedings of the AAAI Conference on Artificial Intelligence, 39(5).DOI: https://doi.org/10.1609/aaai.v39i5.32550
[17] Wang, R., Sun, H., Lin, Y., Zuo, C., Gong, Y., Yin, Y., & Meng, W. (2025). SeqMvRL: A Sequential Fusion Framework for Multi-view Representation Learning. Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 25822-25831.
[18] Dao, T., Fu, D., Ermon, S., Rudra, A., & Ré, C. (2022). FlashAttention: Fast and Memory - Efficient Exact Attention with IO - Awareness. In Advances in Neural Information Processing Systems 35 (NeurIPS 2022), Main Conference Track.
[19] Gu, A., & Dao, T. (2023). Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752.
[20] Riquelme, C., Puigcerver, J., Mustafa, B., Neumann, M., Jenatton, R., Susano Pinto, A., Keysers, D., & Houlsby, N. (2021). Scaling Vision with Sparse Mixture of Experts. Advances in Neural Information Processing Systems, 34.
[21] Chen, T., Moreau, T., Jiang, Z., Zheng, L., Yan, E., Shen, H., Cowan, M., Wang, L., Hu, Y., Ceze, L., Guestrin, C., & Krishnamurthy, A. (2018). TVM: An Automated End-to-End Optimizing Compiler for Deep Learning. Proceedings of the USENIX Annual Technical Conference (USENIX ATC 2018).
[22] Schuhmann, C., Beaumont, R., Venu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., Schramowski, P., Kundurthy, S., Crowson, K., Schmidt, L., Kaczmarczyk, R., & Jitsev, J. (2022). LAION-5B: An open large-scale dataset for training next generation image-text models. Advances in Neural Information Processing Systems, 35.
[23] Bain, M., Nagrani, A., Varol, G., & Zisserman, A. (2021). Frozen in Time: A Joint Video and Image Encoder for End-to-End Retrieval. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, 1728-1738.
[24] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual Explanations From Deep Networks via Gradient-Based Localization. Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, 618-626.
[25] Wu, W., Chang, T., Li, X., Yin, Q., & Hu, Y. (2024). Vision-language navigation: A survey and taxonomy. Neural Computing and Applications, 36, 3291-3316.
[26] Liu, H., Xue, W., Chen, Y., Chen, D., Zhao, X., Wang, K., Hou, L., Li, R., & Peng, W. (2024). A Survey on Hallucination in Large Vision-Language Models. arXiv preprint arXiv:2402.00253.
[27] De Sousa Ribeiro, F., Duarte, K., Everett, M., Leontidis, G., & Shah, M. (2024). Object-centric Learning with Capsule Networks: A Survey. ACM Computing Surveys, 56(11), 1-291. https://doi.org/10.1145/3674500
[28] Franchini, G. (2024). GreenNAS: A Green Approach to the Hyperparameters Tuning in Deep Learning. Mathematics, 12(6), 850. https://doi.org/10.3390/math12060850
[29] Xie, W., Li, X.-H., Cao, C. C., & Zhang, N. L. (2023). ViT-CX: Causal Explanation of Vision Transformers. arXiv preprint arXiv:2211.03064.
[30] LeCun, Y. (2022). A Path Towards Autonomous Machine Intelligence (Version 0.9.2).