Dean&Francis
ISSN 2959-6157

CNN-Transformer Hybrid Models for Object Detection: A Comprehensive Review

Lyuyang Gao

Department of Electronics and Computer Engineering, Shenzhen MSU-BIT University, Shenzhen, China
sz8366541@outlook.com

Abstract:
Initially, conventional convolutional neural networks were the primary approach for object detection, a core computer vision task. However, the emergence of Transformer architecture has significantly enhanced detection accuracy and generalization capabilities, playing a pivotal role in advancing intelligent systems across various domains. Recently, the integration of CNN and Transformer architectures has emerged as a key area of investigation for detecting objects. By combining the complementary advantages of CNNs and Transformers, these hybrid architectures enhance accuracy in various object recognition scenarios. This study commences with a concise overview of CNNs and Transformers, critically analyzing their respective advantages and limitations. Subsequently, we conduct a systematic examination of state-of-the-art hybrid architectures and their optimization strategies. Finally, a comprehensive comparison and summary are presented in tabular form to facilitate clear performance evaluation. These approaches are designed to harness CNNs' superiority in local feature extraction while leveraging Transformers' capacity for global context modeling. At the end of the paper, the prospects of hybrid models in object detection and the insights to guide further research have been discussed.

Keywords: CNN-Transformer Hybrid Model; Serial Architecture Fusion Approach; Parallel Architecture Fusion Method

1. Introduction

As a critical computer vision task, object recognition excels at accurately pinpointing and classifying items in visual data, playing a vital role in fields like self-driving vehicles, healthcare imaging, intelligent monitoring systems, and other domains. Traditional approaches, however, rely on handcrafted features and suffer from limited generalization capability and poor adaptability to complex scenarios.

The advent of Transformer architecture has introduced powerful global feature modeling capabilities, effectively capturing long-range dependencies in images and offering novel solutions for multi-scale object detection challenges. This study focuses on the promising applications of hybrid CNN-Transformer models in object detection, while providing insights for future research directions.

We begin with a concise introduction to CNN and Transformer architecture, including their underlying principles, strengths, and limitations. Subsequently, we present the fundamental concepts and design methodologies of CNN-Transformer hybrid models, followed by a systematic analysis and categorization of prevalent hybrid approaches at different architectural levels. Finally, we discuss the potential of CNN-Transformer hybrid models in advancing object detection performance and outline prospective research opportunities.

2. Comprehensive Analysis of Deep Learning Architectures: From CNNs to Transformers and Hybrid Paradigms

2.1 Convolutional Neural Networks (CNN)

Convolutional Neural Networks (CNNs) are a specialized class of deep learning models designed for processing grid-structured data, such as images and speech. Inspired by biological vision mechanisms, CNNs employ three key principles: local receptive fields, weight sharing, and hierarchical feature extraction. CNNs are fundamentally built using three key elements: feature-extracting convolutions, dimensionality-reducing pooling operations, and decision-making dense layers.At the heart of CNNs, convolution operations employ localized filters with shared parameters to effectively capture spatial patterns in input data. The pooling layer reduces feature map dimensionality via down sampling while preserving critical features and enhancing computational efficiency. Finally, the fully connected layer, positioned at the network‘s terminal stage, flattens the multi-dimensional features extracted by preceding layers into a one-dimensional vector and maps them to the output space through learned weights.

The unique architecture of CNNs, characterized by localized receptive fields and multi-level feature learning, has proven highly effective for visual recognition challenges including image categorization and instance localization. Pioneering architectures like AlexNet [1] and ResNet [2] have driven significant technological breakthroughs. However, their heavy reliance on large-scale datasets and substantial computational resources remains a persistent challenge.

2.2 Transformer

Introduced in 2017 by Vaswani et al., the Transformer model employs self-attention mechanisms as its core computational framework within deep learning systems [3]. Initially developed for NLP sequence tasks like translation, this architecture overcame RNN limitations by implementing fully parallel processing across entire sequences.

The core components of the Transformer include the encoder, decoder, and embedding layer. The embedding layer converts discrete tokens into dense vector representations, while the encoder models long-range contextual relationships. The decoder then sequentially produces outputs by leveraging masked self-attention and encoder-decoder attention mechanisms.

The Transformer’s self-attention architecture and parallel processing capabilities have established it as a groundbreaking advancement in artificial intelligence. However, its high computational resource demands and architectural complexity limit its applicability in resource-constrained scenarios.

2.3 Hybrid CNN-Transformer Model

A notable trend in computer vision research involves integrating convolutional neural networks with transformer-based models to leverage their complementary strengths. By ingeniously integrating the strengths of both frameworks, these hybrid models effectively mitigate the limitations of single-model approaches, substantially enhancing feature extraction capabilities and global modeling performance.

Within hybrid architectures, the CNN component plays two critical roles: first, leveraging its powerful local feature extraction ability to provide low-level visual features for the model; second, enhancing generalization through its inherent inductive bias, particularly when processing high-resolution images. The stage-wise design of CNN modules significantly reduces computational overhead and improves overall efficiency. Meanwhile, the Transformer component focuses on establishing global contextual relationships, capturing long-range dependencies among different image regions via self-attention mechanisms.

This innovative architectural fusion offers dual advantages: first, it achieves complementary integration of local detail features and global semantic information, greatly improving image representation; second, through well-designed structures, it significantly reduces computational complexity while maintaining performance. Currently, such hybrid models have demonstrated outstanding results across various vision tasks and achieved remarkable success in real-world applications.

3. Integration Approaches of CNN and Transformer

3.1 Serial Architecture Fusion Approach

Serial Fusion refers to the sequential connection of CNN and Transformer modules to form a staged feature processing pipeline. The architecture's fundamental principle involves specialized functional partitioning: convolutional networks initially capture localized patterns like basic shapes and surface details through spatial filters, while attention-based modules subsequently model broader contextual relationships across the entire input space. This paper systematically compares and summarizes the models, with the results presented in Table 1.

3.1.1 DETR (Detection Transformer)

Developed by Carion and colleagues, the DETR framework integrates CNN-based feature extraction with transformer-driven sequence analysis in a unified pipeline [4]. The Transformer progressively extracts global contextual information through its encoder-decoder mechanism to generate corresponding object detection anchors. This model achieves superior performance and accuracy in object detection tasks. Figure 1 illustrates the architectural schematic of the DETR (Detection Transformer) model.

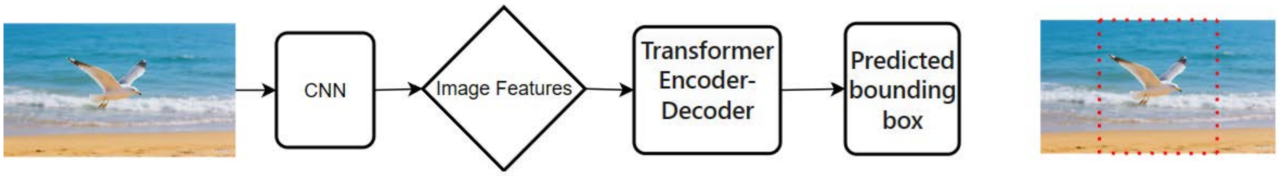

Fig. 1 Structure of DETR model.

3.1.2 Vision Transformer-Faster R-CNN

ViT-FRCNN is a Transformer-based object detection model whose core innovation lies in replacing traditional CNN backbones with a Vision Transformer (ViT) as its feature extraction network[5,6]. The model first processes input images through ViT's patch encoding mechanism, leveraging self-attention to capture global contextual features. The model's output representations undergo spatial reshaping to align with detection requirements, after which they are processed through a cascaded pipeline comprising a region proposal generator and a final detection module. Through end-to-end training, ViT-FRCNN jointly optimizes both the RPN and detection modules, demonstrating superior performance on COCO dataset while exhibiting stronger out-of-domain generalization capabilities and improved large object detection [7]. The key innovation of this approach is its direct utilization of ViT's patch outputs as detection feature maps, thereby eliminating the need for complex multi-scale feature fusion designs commonly employed in conventional detectors.

3.1.3 Convolution-Transformer Network

ConTNet is a hybrid architecture that synergistically models local and global features through alternating stacks of convolutional layers and standard Transformer encoders (STE) [8]. The model initially employs convolutional layers for image down sampling and local feature extraction, followed by serial connections of Transformer encoders to capture long-range dependencies. The final classification or detection output is generated via global pooling and fully connected layers.

The key advantage of ConTNet lies in its dual capability: preserving the inductive biases of convolution (e.g., translation equivariance) while enhancing global context awareness through Transformer modules. On ImageNet classification tasks , ConTNet achieves superior accuracy with lower computational complexity compared to pure Transformer-based models [9]. Notably, when deployed as a backbone for downstream tasks, it significantly outperforms ResNet , particularly in dense prediction tasks such as segmentation, owing to its expanded receptive field enabled by the STE modules.

Table 1. Comparison of Serial-Fusion Hybrid CNN-Transformer Architectures

	Evaluation Index					
Model	Parameter Count(M)	Computation		Strength	Limitation	Ref
DETR	41.0	86 Flops(G)	AP 42%(@ COCO)	An end-to-end detection framework that obviates the requirement for non-maximum suppression (NMS) post-processing, thereby simplifying the pipeline.	The convergence rate is relatively slow, and the detection efficacy for small objects remains unsatisfactory.	[4,7]
ViT-FRCNN	-	-	AP 36.6%(@ COCO)	With minimal fine-tuning, the model achieves rapid adaptation to object detection tasks while maintaining robust performance on out-of-domain (OOD) image datasets.	The model demands large-scale pre-training data and exhibits high computational overhead, posing significant resource requirements.	[5,7]
ConTNet	27.0	217.2 Flops(G)	AP 37.9%(@ COCO)	By employing an alternating stacking strategy, the framework enables stable training convergence without sophisticated data augmentation techniques, while significantly decreasing computational demands.	This architectural design potentially incurs non-negligible computational costs while exhibiting suboptimal generalization performance when applied to limited-scale training data.	[7,8]

3.2 Parallel Architecture Fusion Method

The essence of parallel architecture lies in decoupling the conflict between local perception and global modeling through spatial parallelization and feature fusion. Its core principle is to preserve the spatial detail advantages of CNNs while overcoming the sequence length limitation of Transformers, thereby establishing a new paradigm for high-resolution image understanding. This paper systematically compares and summarizes the models, with the results presented in Table 2.

3.2.1 Conformer-S Model

The Conformer-S model achieves high performance and accuracy in object detection through its innovative parallel CNN-Transformer hybrid architecture and Feature Coupling Unit (FCU) [10]. Key innovations involve bridging localized and holistic representations through Feature Coupling Units (FCUs), simultaneously improving model stability and computational performance. Future directions involve extension to multimodal tasks or integration with compression techniques for further optimization. In summary, Conformer-S represents the cutting edge of CNN-Transformer hybrid models, establishing a new benchmark in object detection research. Figure 2 illustrates the architecture of the Conformer (Convolution-augmented Transformer) model.

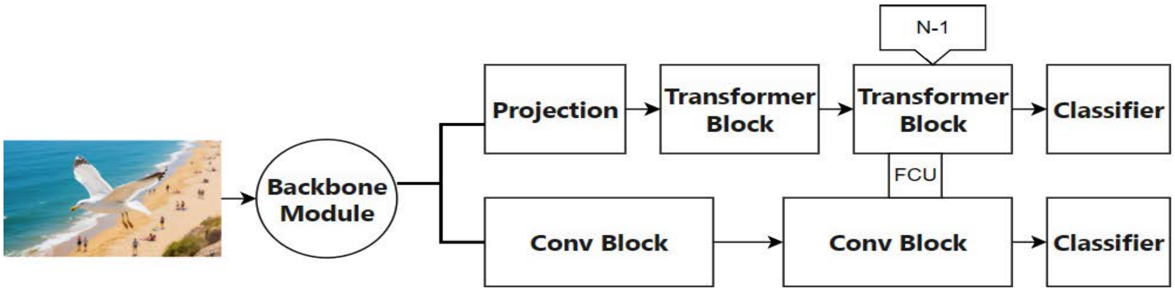

Fig. 2 Structure of Conformer model.

3.2.2 TransMobileNet-Transformer

The TransFusionNet framework, introduced by Wang et al., utilizes a dual-path design to hierarchically capture and integrate holistic image features via encoder-decoder networks, enabling accurate delineation of hepatic lesions and vascular structures [11]. The innovation lies in its simultaneous incorporation of both a Transformer-based global feature extraction encoder and a CNN-based local residual network encoder. These complementary components collectively capture semantic and spatial features from CT images. The model further introduces specialized modules for fusing Transformer and CNN features, along with an edge extraction module, effectively leveraging CNN‘s advantages in local feature extraction and Transformer‘s strengths in global context modeling. This design significantly enhances both accuracy and robustness in medical image segmentation.

3.2.3 MobileNet-Transformer

Mobile-Former jointly developed by Microsoft and the Chinese Academy of Sciences in 2021, represents a lightweight hybrid architecture that parallelly integrates MobileNet with Transformer [12]. A pivotal advancement lies in its Bidirectional Cross-Attention Bridge, enabling seamless integration of convolutional networks' region-specific feature learning with Transformers' long-range dependency modeling. This architecture achieves an optimal balance between computational efficiency and strong representational power [13].

Table 2. Comparison of Parallel-Fusion Hybrid CNN-Transformer Architectures

	Evaluation Index					
Model	Parameter Count(M)	Computation		Strength	Limitation	Ref
Conformer-S	89	162 Flops(G)	AP 46.6%(@ COCO)	The proposed feature coupling module (FCM) effectively fuses localized feature patterns with holistic contextual representations, thereby strengthening feature discriminability through cross-scale interaction.	The proposed architecture demonstrates non-trivial computational complexity, which consequently imposes considerable difficulties in model training convergence.	[7,10]
TransFusionNet	7.96	-	Latency(ms) 114.9 (On the hardware platform comprising an NVIDIA Titan V100 GPU and Intel Core i7 CPU)	The model achieves high-precision segmentation and multi-task collaboration, and enables embedded deployment after quantization.	The boundary target segmentation exhibits inadequate stability, elevated inter-module coupling, and intricate parameter optimization.	[11,14]
Mobile-Former	8.4	9.8 Madds(G)	AP 38.0%(@ COCO)	The approach significantly pushes the accuracy limits of efficient architectures while operating within strict computational constraints.	Caution should be exercised when evaluating in scenarios involving extreme lightweight requirements or limited data availability.	[7,12]

4. Discussion and Future work

The ongoing advancement of sophisticated AI systems is accelerating progress in machine intelligence, delivering groundbreaking results in both linguistic understanding and visual interpretation - heralding a fundamental shift in how society operates and interacts. The deep integration of hybrid models with visual foundation models has emerged as a critical research challenge to be addressed.

First, in certain specialized image processing scenarios, researchers frequently encounter challenges including insufficient data samples, labor-intensive annotation processes, and complex semantic interpretation. Looking ahead, we can leverage the inherent advantages of foundation models to facilitate image data generation, annotation, and interpretation. Subsequently, through integration with hybrid models, we can conduct in-depth analysis of image data to extract multi-dimensional features, thereby establishing a robust foundation for subsequent task evaluation.

Second, multimodal learning is emerging as a pivotal trend for future development. The development of novel hybrid foundation models(e.g., Swin-Transformer) that combine CNN and Transformer architectures will enable the processing of multimodal data (encompassing text, images, audio, etc.), achieving cross-modal interaction and uncovering the correlations and complementarity between different modalities [15]. Such hybrid models are expected to provide substantial momentum for the continued research and widespread adoption of visual foundation models.

References

[1] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems, 2012, 25: 1097-1105.

[2] Fangfang. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 770-778.

[3] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. Advances in Neural Information Processing Systems, 2017, 30: 5998-6008.

[4] Carion N, Massa F, Synnaeve G, et al. End-to-end object detection with transformers. European Conference on Computer Vision, 2020: 213-229.

[5] Beal J, Kim E, Tzeng E, et al. Toward transformer-based object detection. arXiv preprint arXiv:2012.09958, 2020.

[6] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale. International Conference on Learning Representations, 2021.

[7] Lin T Y, Maire M, Belongie S, et al. Microsoft COCO: Common objects in context. European Conference on Computer Vision, 2014: 740-755.

[8] Yan S, Xiong X, Arnab A, et al. ConTNet: Why not use convolution and transformer at the same time? arXiv preprint arXiv:2104.13497, 2021.

[9] Deng J, Dong W, Socher R, et al. ImageNet: A large-scale hierarchical image database. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2009: 248-255.

[10] Peng Z, Huang W, Gu S, et al. ConFormer: Local features coupling global representations for visual recognition. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021: 367-376.

[11] Wang X, Zhang X D, Wang G, et al. TransFusionNet: Semantic and spatial features fusion framework for liver tumor and vessel segmentation under Jetson TX2. IEEE Transactions on Medical Imaging, 2022, 41(5): 1123-1135.

[12] Chen Y, Dai X, Chen D, et al. Mobile-Former: Bridging MobileNet and transformer. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 5270-5279.

[13] Zhu X, Su W, Lu L, et al. Deformable DETR: Deformable transformers for end-to-end object detection. International Conference on Learning Representations, 2021.

[14] Bai X, Hu Z, Zhu X, et al. TransFusion: Robust LiDAR-camera fusion for 3D object detection with transformers. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 1718-1727.

[15] Liu Ze, Lin Yutong, Cao Yue, Hu Han, et al. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. arXiv:2103.14030, 2021.