

# A REVIEW OF DEVELOPMENT OF A HIERARCHICAL ATTENTION-GUIDED DEEP CONVOLUTIONAL NETWORK FOR CONTEXT-AWARE IMAGE UNDERSTANDING WITH DYNAMIC FEATURE SUPPRESSION IN PYTHON

Km. Mahima Verma<sup>1</sup>, Mrs. Arifa Khan<sup>2</sup>

<sup>1</sup>Master of Technology, Computer Science and Engineering, Lucknow Institute of Technology, Lucknow, India

<sup>2</sup>Assistant Professor, Department of Computer Science and Engineering, Lucknow Institute of Technology, Lucknow, India

\*\*\*

**Abstract** - The rapid evolution of deep convolutional neural networks (CNNs) has significantly advanced image understanding. However, traditional models often struggle to capture long-range dependencies and salient contextual information, treating all spatial and channel-wise features uniformly. This uniform processing leads to computational inefficiency and suboptimal performance in complex scenes where context is key. The emergence of hierarchical attention mechanisms and dynamic feature suppression offers a promising paradigm to address these limitations. This paper presents a systematic review of attention-guided deep learning architectures designed for context-aware image understanding. We survey the landscape from soft and hard attention to modern Transformers and dynamic gating mechanisms. We synthesize the literature into a taxonomy, discuss the theoretical underpinnings of feature suppression, and analyze the integration of these components into hierarchical networks. We identify a critical research gap in the joint optimization of attention for both selection (what to look at) and suppression (what to ignore). The review concludes by outlining open challenges and proposing future research directions, including the development of unified frameworks and efficient Python-based implementations for real-world applications.

**Key Words:** Attention Mechanisms; Context-Aware Image Understanding; Deep Learning; Dynamic Feature Suppression; Hierarchical Neural Networks; Computer Vision

## 1. INTRODUCTION

### 1.1. Background and Motivation

The field of computer vision has witnessed a remarkable evolution over the past decade, transitioning from basic image classification tasks to complex scene understanding problems such as semantic segmentation, image captioning, and visual question answering. This journey began with groundbreaking work on large-scale image classification using deep convolutional neural networks (Krizhevsky et al., 2012), which demonstrated that hierarchical feature learning could achieve unprecedented accuracy on datasets like ImageNet. Subsequent advances in network architecture, including the introduction of VGG (Simonyan

and Zisserman, 2015) and residual learning in ResNet (He et al., 2016), progressively pushed the boundaries of what deep models could accomplish. These foundational developments paved the way for more sophisticated vision tasks that require not merely object recognition but a holistic understanding of visual scenes, including relationships between objects, spatial configurations, and semantic context (Long et al., 2015; Ren et al., 2017).

Despite these advances, standard convolutional neural networks possess inherent limitations that constrain their ability to achieve genuine context-aware understanding. Traditional CNNs operate with fixed receptive fields determined by kernel sizes and network depth, which fundamentally limits their capacity to capture long-range dependencies and global contextual information (Wang et al., 2018). While stacking multiple convolutional layers can theoretically expand the receptive field, in practice, the effective receptive field is often much smaller than the theoretical maximum due to the concentration of influence in central regions (Luo et al., 2016). Furthermore, conventional CNNs process all spatial locations and feature channels uniformly, treating every region of an image with equal importance regardless of its relevance to the task at hand. This uniform processing paradigm leads to two significant shortcomings: computational inefficiency, as resources are expended on processing irrelevant background regions, and suboptimal performance in complex scenes where contextual relationships are crucial for accurate interpretation (Hu et al., 2018).

The "context-awareness" problem in computer vision fundamentally concerns the challenge of understanding visual elements not in isolation but in relation to their surroundings. Recognising an object often requires understanding the scene in which it appears—a small, cylindrical object might be identified as a cup when situated on a dining table but interpreted differently when found in a bathroom setting (Oliva and Torralba, 2007). Similarly, interpreting human actions requires understanding the objects involved and the environment where the action occurs. This contextual reasoning, which comes naturally to human perception, proves remarkably challenging for artificial vision systems. Early approaches to incorporating context involved multi-scale architectures and spatial

pyramid pooling (Lazebnik et al., 2006; He et al., 2015), which aggregated information at multiple resolutions, and atrous convolution techniques that expanded receptive fields without increasing parameters (Chen et al., 2017). While these methods represented important steps toward context-aware processing, they lacked the adaptive, selective mechanisms characteristic of human visual attention.

### 1.1.1. The Evolution from Classification to Scene Understanding

The progression from image classification to comprehensive scene understanding marks a fundamental shift in the objectives of computer vision systems. Image classification tasks require assigning a single label to an entire image, a problem that can often be solved by identifying the most salient object present (Russakovsky et al., 2015). In contrast, scene understanding demands detailed interpretation of all elements within an image, their relationships, and the overall semantic context. Semantic segmentation requires pixel-level classification, where each pixel must be assigned to a category, necessitating both local detail preservation and global contextual coherence (Long et al., 2015). Image captioning goes further, requiring the generation of natural language descriptions that capture not only the objects present but also their attributes, actions, and interactions (Vinyals et al., 2015). Visual question answering challenges models to reason about images based on natural language questions, often requiring sophisticated reasoning about spatial relationships, counting, and commonsense knowledge (Antol et al., 2015). Each of these tasks demands forms of contextual understanding that exceed the capabilities of standard CNNs operating with fixed receptive fields and uniform feature processing, motivating the development of more sophisticated attention-guided architectures.

## 1.2. The Rise of Attention and Dynamic Mechanisms

The concept of attention in neural networks draws direct inspiration from the human cognitive system, which possesses a remarkable ability to focus processing resources on salient regions of the visual field while suppressing irrelevant information (Itti et al., 1998). In human perception, attention operates through a combination of bottom-up, saliency-driven mechanisms and top-down, task-guided selection, enabling efficient processing of complex visual scenes despite limited neural resources. This biological paradigm has proven highly influential in artificial intelligence, leading to the development of computational attention mechanisms that enable deep networks to selectively focus on informative features while ignoring distractors (Hassanin et al., 2024).

The history of attention mechanisms in neural networks traces an interesting trajectory from vision to natural language processing and back to vision. Early computational models of visual attention, such as the saliency-based system

proposed by Itti et al. (1998), demonstrated that biologically-inspired attention could efficiently identify important regions in complex scenes. This work established foundational principles of computational attention that would later influence deep learning approaches. A significant milestone occurred when Mnih et al. (2014) introduced the recurrent attention model (RAM), which first incorporated attention mechanisms into recurrent neural networks for image classification. This work demonstrated that attention could enable networks to process images sequentially, focusing on different regions over time, much like human eye movements.

The transformative impact of attention in deep learning, however, is most frequently associated with its application in natural language processing. Bahdanau et al. (2015) introduced attention mechanisms to neural machine translation, enabling the model to align target words with relevant source words during translation. This work addressed the fundamental limitation of fixed-length context vectors in encoder-decoder architectures and demonstrated that attention could significantly improve performance on sequence-to-sequence tasks. The subsequent introduction of the Transformer architecture (Vaswani et al., 2017), which dispensed with recurrent and convolutional processing entirely in favour of pure self-attention, revolutionised natural language processing and established attention as a fundamental building block of modern deep learning.

The success of attention in NLP catalysed its reintroduction to computer vision with renewed vigour. The squeeze-and-excitation network (SENet) (Hu et al., 2018) introduced channel attention, enabling networks to adaptively recalibrate channel-wise feature responses by modelling interdependencies between channels. This simple yet effective mechanism could be incorporated into any CNN architecture and yielded consistent performance improvements across multiple tasks. The convolutional block attention module (CBAM) (Woo et al., 2018) extended this idea by sequentially applying channel and spatial attention, demonstrating that complementary attention mechanisms could work synergistically. Non-local neural networks (Wang et al., 2018) brought self-attention to vision by enabling each position to attend to all other positions, capturing long-range dependencies that traditional convolutions could not. Finally, the Vision Transformer (ViT) (Dosovitskiy et al., 2021) applied pure transformer architectures to image classification, demonstrating that with sufficient data, attention-based models could outperform convolutional networks on vision tasks. This trajectory of development has established attention as an indispensable component of modern computer vision architectures.

### 1.2.1. Defining Dynamic Feature Suppression

Dynamic feature suppression represents a distinct yet complementary concept to attention, concerned specifically

with the active inhibition of irrelevant or noisy features rather than merely weighting them less heavily. While attention mechanisms typically compute importance weights that scale feature responses, suppression mechanisms can entirely eliminate or gate the flow of information from unimportant channels or spatial locations (Gao et al., 2019). This distinction carries important implications for both model performance and computational efficiency. The importance of active suppression arises from the observation that the saliency of neurons in convolutional layers is highly input-dependent—a channel that proves critical for one image may be nearly irrelevant for another (Gao et al., 2019). Static pruning methods that permanently remove channels based on average importance across a dataset inevitably sacrifice the capability to handle inputs for which those pruned channels would have been essential. Dynamic suppression preserves the full network structure while skipping computation for unimportant channels at run-time, maintaining model capacity while reducing computational cost.

### 1.3. Scope and Contributions of this Review

This paper presents a systematic review of attention-guided deep learning architectures for context-aware image understanding, with a specific focus on hierarchical networks and dynamic feature suppression mechanisms. It is important to state explicitly that this work constitutes a review paper rather than a novel methodological contribution. Our objective is to synthesise the extensive and rapidly growing literature on attention mechanisms in computer vision, identify patterns and principles that transcend individual contributions, and provide a structured analysis that can guide future research in this area. The review adopts a critical perspective, evaluating not only the achievements of existing work but also identifying limitations, open challenges, and promising directions for further investigation.

The scope of this review is carefully delimited to ensure focused and coherent analysis. We concentrate on hierarchical, multi-scale attention mechanisms that operate at multiple levels of feature abstraction, from low-level edges and textures to high-level semantic concepts. This hierarchical perspective is essential for context-aware understanding, as contextual information manifests at multiple scales—local context informs object boundaries, while global context determines scene semantics (Chen et al., 2018). We specifically examine attention-guided deep networks designed for context-aware image understanding, including applications in semantic segmentation, scene parsing, image captioning, and visual question answering. A distinctive focus of this review is the examination of dynamic feature suppression mechanisms, which have received less systematic attention in previous surveys despite their growing importance for both efficiency and robustness. We consider suppression mechanisms including channel gating,

spatial pruning, and dynamic convolution, analysing their relationship to attention and their role in comprehensive context-aware systems.

#### 1.3.1. Primary Contributions of This Review

This review makes four primary contributions to the literature on attention mechanisms in computer vision. First, we develop a novel taxonomy of attention-suppression mechanisms that organises the literature according to fundamental design dimensions rather than superficial architectural differences. This taxonomy distinguishes between soft and hard attention, channel, spatial, and mixed attention, and further categorises suppression mechanisms according to their granularity (channel-wise, spatial-wise, or element-wise) and their dynamic or static nature. By providing a structured framework for understanding the relationships between different approaches, this taxonomy aims to facilitate comparison and guide architectural design decisions.

Second, we provide a synthesis of architectural design patterns that recur across successful attention-guided networks. Rather than presenting an exhaustive catalogue of individual models, we identify common building blocks and design principles—such as the encoder-decoder paradigm with attention in skip connections, the integration of self-attention for long-range dependency modelling, and the combination of complementary attention mechanisms. This synthesis aims to distil practical insights that can inform the development of new architectures for context-aware understanding.

## 2. LITERATURE REVIEW

### 2.1. Foundational Concepts: Context in Computer Vision

The concept of context has long been recognised as fundamental to visual perception, both in biological and artificial vision systems. In computer vision, context refers to the information surrounding a target element that aids in its interpretation, encompassing everything from local pixel neighbourhoods to global scene semantics (Oliva and Torralba, 2007). The integration of contextual information addresses a fundamental limitation of isolated object recognition: many objects are inherently ambiguous when viewed in isolation and only become identifiable through their relationship with the surrounding environment. For instance, a small cylindrical object might be interpreted as a cup when situated on a dining table, but as a toothbrush holder when found in a bathroom setting. This contextual disambiguation is essential for robust image understanding systems that must operate in unconstrained real-world environments.

### 2.1.1. Global versus Local Context

Early approaches to incorporating context in deep neural networks focused on multi-scale architectures that could simultaneously capture both fine-grained local details and coarse global scene structure. The fundamental insight underlying these approaches is that different semantic information manifests at different spatial scales: object boundaries require high-resolution local context, while scene category and overall spatial layout are best captured at lower resolutions with larger receptive fields (Farabet et al., 2013).

Spatial pyramid pooling represented one of the earliest systematic attempts to address multi-scale context aggregation. Lazebnik et al. (2006) introduced the concept of pyramid matching in the context of scene recognition, partitioning images into increasingly fine subregions and computing histograms of local features within each subregion. This approach was later adapted for deep learning by He et al. (2015), who proposed spatial pyramid pooling in convolutional neural networks to generate fixed-length representations regardless of input size, enabling the network to aggregate information at multiple scales simultaneously.

A parallel line of development emerged from the recognition that standard convolutional operations with fixed kernels could not efficiently capture context at multiple scales without incurring prohibitive computational costs. Atrous convolution, also known as dilated convolution, provided an elegant solution to this problem by introducing a dilation rate parameter that controls the spacing between kernel weights (Chen et al., 2018). This mechanism allows filters to operate with enlarged receptive fields without increasing parameter count or computational complexity, effectively enabling the network to capture multi-scale context through parallel branches with different dilation rates. As Chen et al. (2018) demonstrated, atrous convolution proves particularly effective for dense prediction tasks where maintaining spatial resolution while expanding receptive fields is critical.

### 2.1.2. Context via Recurrent Models

While multi-scale convolutional approaches effectively capture spatial context at multiple resolutions, they process all image regions simultaneously and uniformly. An alternative paradigm draws inspiration from human visual perception, which sequentially samples the visual environment through saccadic eye movements, focusing attention on salient regions while maintaining a global scene representation. Recurrent neural networks (RNNs) and their variants, particularly Long Short-Term Memory (LSTM) networks, provided a natural framework for implementing such sequential processing in deep learning systems.

The recurrent attention model (RAM) introduced by Mnih et al. (2014) represented a seminal contribution to this direction. RAM processes images by sequentially selecting and processing small image regions, building up a representation over time through a recurrent network that learns both where to look and how to interpret the gathered information. This approach demonstrated that sequential attention could achieve competitive classification performance while processing only a fraction of the image pixels, highlighting the computational efficiency benefits of selective contextual sampling.

## 2.2. Attention Mechanisms in Vision: A Taxonomy

The success of attention mechanisms in neural machine translation (Bahdanau et al., 2015) catalysed extensive research into attention-based models for computer vision, yielding a diverse landscape of mechanisms that operate at different levels of representation and employ various computational strategies. Organising this literature into a coherent taxonomy requires considering multiple dimensions: the domain of application (spatial, channel, or hybrid), the nature of attention computation (soft deterministic or hard stochastic), and the architectural level at which attention operates (single-scale or hierarchical). This section develops such a taxonomy to provide a structured framework for understanding the relationships between different attention approaches.

### 2.2.1. Soft versus Hard Attention

The distinction between soft and hard attention centres on the differentiability and stochasticity of the attention mechanism, with profound implications for training methodology and architectural design. Soft attention mechanisms compute a weighted combination of all input elements, with weights typically derived from a differentiable compatibility function, enabling end-to-end training via standard backpropagation. Hard attention mechanisms, in contrast, select a subset of input elements through stochastic sampling, rendering the selection process non-differentiable and necessitating alternative training approaches such as reinforcement learning.

The Selective Kernel Network (SKNet) (Li et al., 2019) extended channel attention by introducing dynamic kernel selection based on attention mechanisms. SKNet employs multiple branches with different kernel sizes, aggregates information across branches through element-wise summation, and then applies attention mechanisms to selectively emphasise features from different kernel scales based on input content. This enables the network to adaptively adjust its receptive field size depending on the scale of objects present in the input, effectively learning to select appropriate convolutional kernel sizes through soft attention.

### 2.2.2. Hierarchical and Multi-scale Attention

While single-scale attention mechanisms can effectively highlight important features at a fixed resolution, natural images contain structures at multiple scales that require attention operating at corresponding levels of abstraction. Hierarchical attention mechanisms address this requirement by applying attention at different levels of feature hierarchies, from low-level edge and texture representations in early network layers to high-level semantic representations in deeper layers.

Pyramid Attention Networks (PAN) introduced by Li et al. (2018) explicitly address multi-scale attention through a pyramidal structure that applies attention at multiple feature resolutions. PAN first generates attention maps at multiple scales, then uses these maps to weight feature maps at corresponding scales, and finally fuses the weighted multi-scale features through upsampling and summation. This design enables the network to capture both fine-grained details through high-resolution attention and broad contextual relationships through low-resolution attention, with information flowing both bottom-up and top-down through the attention pyramid.

### 2.2.3. Self-Attention and Transformers

The introduction of the Transformer architecture (Vaswani et al., 2017) in natural language processing revolutionised sequence modelling by demonstrating that pure self-attention, without recurrent or convolutional components, could achieve state-of-the-art performance while offering superior parallelisation and the ability to capture long-range dependencies. This success catalysed extensive research into adapting Transformer architectures for computer vision, fundamentally reshaping the landscape of visual representation learning.

## 2.3. Dynamic Feature Suppression: Gating and Selection

While attention mechanisms focus on amplifying important features, an equally important aspect of selective information processing is the suppression of irrelevant or distracting information. Dynamic feature suppression encompasses mechanisms that actively gate or prune features based on input content, preventing irrelevant information from propagating through the network and potentially corrupting representations. This section reviews approaches to dynamic suppression at different levels of representation, drawing inspiration from both computational efficiency considerations and biological principles of selective attention.

The biological inspiration for dynamic suppression derives from neurocognitive research demonstrating that the brain actively suppresses distracting information rather than merely failing to attend to it. Di Bello et al. (2021) provide a

unified neurocognitive framework describing how the prefrontal cortex implements both proactive suppression—strategically prioritising task-relevant information at the expense of irrelevant information—and reactive suppression—interrupting the ongoing processing of distractors that have captured attention. This dual-mechanism perspective highlights that effective selective processing requires both preventing distraction before it occurs and terminating distraction when it breaks through initial filters.

### 2.3.1. Channel-wise Suppression

Channel-wise suppression mechanisms operate along the channel dimension of convolutional feature maps, selectively inhibiting or gating entire feature channels based on their estimated relevance to the current input. The Squeeze-and-Excitation Network (SENet) (Hu et al., 2018) can be interpreted as implementing a form of soft channel suppression: the excitation operation produces channel weights that can approach zero for unimportant channels, effectively suppressing their contribution to subsequent layers. However, SENet's suppression is multiplicative and continuous—channels with low weights contribute minimally but are not completely eliminated, and the computation required to process these channels remains unchanged.

More aggressive channel-wise suppression mechanisms explicitly zero out or skip computation for unimportant channels, achieving both representational benefits and computational savings. Gao et al. (2019) introduced GaterNet, a dual-path architecture where a lightweight "gater" network learns to generate binary gates that selectively activate or deactivate channels in a main "backbone" network. The gater network processes the same input as the backbone but with reduced capacity, learning to predict which channels will be useful for the current input. Channels with zero gates are completely skipped during backbone computation, reducing both the number of operations and the memory footprint. This dynamic channel gating enables the network to adapt its effective capacity to each input, allocating more computation to challenging examples while processing simple examples efficiently.

### 2.3.2. Spatial-wise Suppression

Spatial-wise suppression mechanisms operate on the spatial dimensions of feature maps, selectively inhibiting or pruning responses at specific spatial locations. This form of suppression recognises that in natural images, many spatial regions contain background or irrelevant content that can be safely ignored without compromising task performance. By suppressing such regions, spatial attention mechanisms can focus computational resources on informative areas while reducing the influence of distracting background patterns.

The connection between spatial suppression and sparsity is fundamental: suppressing spatial locations effectively introduces spatial sparsity in feature representations, with only a subset of locations propagating significant information to subsequent layers. This sparsity can be exploited for computational efficiency through sparse convolution operations that only process active locations (Graham et al., 2018), though practical implementations often require specialised hardware or software support for irregular computation patterns.

### 2.3.3. Dynamic Convolution

Dynamic convolution represents a more fundamental form of feature adaptation than attention or gating, where the convolutional kernels themselves are generated or assembled dynamically based on input content. Rather than selecting which features to emphasise or suppress within a fixed kernel representation, dynamic convolution adapts the kernels to the input, enabling content-parametric filtering that can capture input-specific patterns.

CondConv (Yang et al., 2019) introduces conditionally parameterised convolutions where kernel weights are computed as a linear combination of multiple expert kernels, with combination weights predicted from the input. Specifically, for each input example, CondConv computes routing weights through a lightweight gating function, then combines multiple convolutional kernels by weighted summation to produce an input-specific kernel. This approach increases model capacity without increasing inference cost proportionally, as the kernel combination requires only  $O(K)O(K)$  operations where  $K$  is the number of experts, while the subsequent convolution uses the combined kernel of standard size. Yang et al. (2019) demonstrate that CondConv improves accuracy across multiple architectures while maintaining efficient inference, effectively learning to adapt convolutional filters to input content.

## 2.4. Integration into Downstream Tasks: Context-Aware Understanding

The ultimate test of attention and suppression mechanisms lies in their ability to improve performance on downstream tasks requiring genuine context-aware understanding. This section reviews how the techniques discussed above have been integrated into architectures for semantic segmentation, image captioning, scene parsing, and visual question answering, examining both the specific design choices that enable effective context modelling and the empirical benefits demonstrated on standard benchmarks.

### 2.4.1. Semantic Segmentation

Semantic segmentation, the task of assigning a class label to every pixel in an image, has served as a primary testbed for context-aware architectures due to its inherent requirement

for integrating local detail with global scene understanding. Accurate segmentation demands both precise boundary localisation, requiring high-resolution features, and consistent labelling across object instances, requiring contextual reasoning about object relationships and scene semantics.

OCRNet (Object-Contextual Representations) introduced by Yuan et al. (2020) represents a significant advance in context modelling for segmentation by explicitly representing object-contextual information. The key insight underlying OCRNet is that the most relevant context for a pixel comes from other pixels belonging to the same object category, rather than from all pixels indiscriminately. To exploit this, OCRNet first estimates a coarse object region representation through a segmentation head, then computes pixel-to-region relationships that weight the contribution of each object region to each pixel's representation. This approach enables each pixel to attend selectively to regions containing the same object category, effectively implementing a form of object-aware contextual aggregation. Yuan et al. (2020) demonstrate that OCRNet achieves state-of-the-art performance on multiple segmentation benchmarks, confirming the value of object-aware context modelling.

### 2.4.2. Scene Parsing and Image Captioning

Scene parsing, which involves segmenting an image into semantically meaningful regions corresponding to objects and their parts, demands even richer contextual understanding than semantic segmentation alone. The Pyramid Scene Parsing Network (PSPNet) introduced by Zhao et al. (2017) addresses this through pyramid pooling modules that aggregate context at multiple scales. PSPNet applies pooling operations at different grid scales, producing pooled representations that capture global, regional, and local context, then upsamples and concatenates these representations to form a comprehensive contextual feature. This design enables the network to incorporate information from the entire scene when parsing each pixel, reducing the likelihood of local confusion (e.g., mistaking a boat for a car when both appear near water).

Image captioning requires generating natural language descriptions that capture not only the objects present but also their attributes, relationships, and activities. The Up-Down attention model (Anderson et al., 2018) introduced a two-level attention mechanism that first attends to objects (through bottom-up region proposals) and then attends to relationships between objects (through top-down attention conditioned on the current language state). This approach recognises that effective caption generation requires both identifying salient objects and understanding how they relate to each other, with attention operating at both levels. Anderson et al. (2018) demonstrated state-of-the-art performance on multiple captioning benchmarks, establishing the value of hierarchical attention that combines bottom-up saliency with top-down task guidance.

### 2.4.3. Visual Question Answering

Visual Question Answering (VQA) presents perhaps the most demanding test of context-aware understanding, requiring systems to answer natural language questions about images by integrating visual perception, language understanding, and commonsense reasoning. Questions can range from simple object identification ("What colour is the car?") to complex relational reasoning ("How many children are standing to the left of the woman holding an umbrella?"), each requiring different forms of visual attention and contextual aggregation.

Stacked Attention Networks (SAN) introduced by Yang et al. (2016) address VQA through multiple attention layers that progressively refine the focus on image regions relevant to the question. The first attention layer identifies coarse regions potentially containing relevant information, while subsequent layers refine this focus based on the question context and previously attended regions. This multi-step attention process mirrors human visual search behaviour, where initial glances identify promising regions followed by closer inspection of detailed content. Yang et al. (2016) demonstrated that stacked attention significantly improves VQA accuracy, particularly for questions requiring multi-step reasoning.

## 3. ANALYSIS OF ARCHITECTURAL DESIGN PATTERNS

### 3.1. The Encoder-Decoder Paradigm

The encoder-decoder architecture has emerged as a dominant paradigm for dense prediction tasks in computer vision, including semantic segmentation, depth estimation, and image-to-image translation. This design pattern consists of an encoder network that progressively reduces spatial resolution while increasing feature dimensionality, capturing increasingly abstract representations, and a decoder network that progressively recovers spatial resolution to produce output at the original input dimensions (Badrinarayanan et al., 2017). The encoder-decoder structure naturally supports multi-scale feature learning and has proven particularly effective when combined with attention mechanisms that selectively emphasise relevant information at different stages of the processing pipeline.

The integration of attention modules within encoder-decoder architectures follows several established patterns, each offering distinct advantages for context modelling. The choice of insertion points reflects fundamental trade-offs between computational efficiency, representational capacity, and the nature of contextual information being captured. Understanding these design patterns provides insight into how attention mechanisms can be most effectively deployed for context-aware image understanding.

### 3.1.1. Attention in the Bottleneck

The bottleneck represents the lowest spatial resolution and highest semantic abstraction point in the encoder-decoder pipeline, typically occurring at the transition between encoding and decoding stages. Inserting attention mechanisms at the bottleneck enables the network to model global context and long-range dependencies before spatial information is progressively restored during decoding. This design choice is particularly effective for tasks requiring understanding of overall scene structure and relationships between distant image regions.

The non-local neural network proposed by Wang et al. (2018) exemplifies attention at the bottleneck, inserting non-local blocks at intermediate network stages where feature maps have relatively low spatial resolution but high channel dimensionality. At this resolution, the quadratic complexity of self-attention becomes computationally tractable, enabling the modelling of relationships between all spatial positions. The non-local block computes pairwise affinities across the entire spatial extent, generating context-aggregated features that incorporate information from all image regions. When placed at the bottleneck, this global contextual information can then be propagated to higher resolutions through the decoder, ensuring that fine-grained predictions benefit from holistic scene understanding.

### 3.1.2. Attention in Skip Connections

Skip connections, which directly transmit high-resolution features from encoder layers to corresponding decoder layers, provide an alternative and complementary location for attention insertion. The U-Net architecture (Ronneberger et al., 2015) popularised this design, demonstrating that combining high-resolution encoder features with upsampled decoder features through concatenation substantially improves segmentation accuracy, particularly along object boundaries where fine detail is essential.

The combination of bottleneck and skip connection attention represents a particularly powerful design pattern, enabling networks to capture both global context through low-resolution attention and fine detail through high-resolution attention. This dual-attention approach characterises state-of-the-art architectures for dense prediction tasks, with OCRNet (Yuan et al., 2020) exemplifying the integration of object-region attention at multiple scales to achieve both contextual awareness and precise localisation.

## 3.2. Complexity Analysis

The practical deployment of attention mechanisms in deep learning systems requires careful consideration of computational and memory complexity. Different attention designs exhibit vastly different scaling behaviours with respect to input resolution and feature dimensionality, with important implications for both training feasibility and

inference efficiency. This section analyses the complexity characteristics of major attention families and discusses their practical implications for context-aware image understanding systems.

### 3.2.1. Computational Complexity (FLOPs)

Floating-point operations (FLOPs) provide a hardware-independent measure of computational cost that enables comparison across different attention mechanisms. The complexity analysis reveals fundamental trade-offs between representational power and computational efficiency that guide architectural design choices.

## 4. CHALLENGES IN IMPLEMENTATION

Despite the maturity of deep learning frameworks and the richness of the Python ecosystem, implementing attention-guided networks with dynamic feature suppression presents several significant challenges. These challenges span the entire development lifecycle, from model design and training to deployment and optimisation, and require careful consideration to achieve both functional correctness and practical efficiency.

### 4.1. Difficulty in Implementing Hard Attention

The implementation of hard attention mechanisms poses fundamental difficulties arising from the non-differentiability of discrete selection operations. Unlike soft attention, which computes continuous weights that can be propagated through standard backpropagation, hard attention makes discrete decisions about which features to process—for instance, selecting a subset of spatial locations or binary gating of channels. These discrete decisions create a discontinuity in the computational graph, preventing the direct application of gradient-based optimisation.

As articulated in research on quantum hard attention mechanisms, the non-differentiability challenge has constrained the widespread applicability of hard attention in deep learning. Various strategies have been developed to circumvent this limitation, each with its own trade-offs. The straight-through estimator, popularised by Bengio et al. (2013), approximates the gradient of discrete thresholding operations by treating them as identity functions during backward propagation. This approach enables end-to-end training but introduces bias in gradient estimation that can affect convergence. The Gumbel softmax relaxation provides a differentiable approximation to discrete sampling by replacing the argmax operation with a softmax over Gumbel-perturbed logits, with a temperature parameter controlling the sharpness of the approximation. During training, the temperature can be annealed to approach discrete decisions at inference time while maintaining differentiability throughout optimisation.

### 4.2. Training Stability of Transformers on Small Datasets

Vision Transformers and their variants present significant training challenges when applied to small or medium-sized datasets, stemming from their reduced inductive bias compared to convolutional neural networks. As documented in recent CVPR workshop proceedings, ViTs typically require substantially larger training datasets to learn local feature representations effectively, with performance on ImageNet-1K-scale datasets often falling short of comparably-sized CNNs without extensive data augmentation or pre-training.

The fundamental issue lies in the self-attention mechanism's flexibility: while this flexibility enables the modelling of complex relationships, it also means the network must learn spatial locality and translation equivariance from data rather than having these properties built into the architecture. On small datasets, the statistical evidence for these inductive biases is insufficient, leading to overfitting and poor generalisation. The Multi-Gradient Image Transformer (MGiT) approach proposed by researchers addresses this challenge through parallel training with a compact auxiliary ViT that adaptively optimises the target network's weights, demonstrating that specialised training strategies can partially compensate for limited data.

## 5. EVALUATION, DATASETS, AND BENCHMARKS

### 5.1. Standard Datasets

The evaluation of attention-guided deep learning models for context-aware image understanding relies on a collection of standardised datasets that have become established benchmarks within the computer vision community. These datasets are carefully curated to encompass diverse visual scenes, rich annotations, and tasks that require genuine contextual reasoning, thereby providing meaningful grounds for comparing different architectural approaches.

The Common Objects in Context (COCO) dataset represents one of the most widely adopted benchmarks for context-aware image understanding, encompassing object detection, segmentation, and captioning tasks. As González-Chávez et al. (2023) note, COCO has become a cornerstone dataset for image captioning research, providing complex scenes with multiple interacting objects that necessitate contextual reasoning for accurate description. The dataset contains over 200,000 images with detailed annotations including object instance segmentation, stuff segmentation, and five human-written captions per image, enabling comprehensive evaluation of models' ability to understand both objects and their contextual relationships. The COCO-Stuff variant extends this with additional stuff category annotations, providing even richer contextual information for dense prediction tasks (You et al., 2025).

## 5.2. Evaluation Metrics

The assessment of attention-guided models for context-aware understanding employs a multifaceted suite of evaluation metrics, each designed to capture different dimensions of model performance. These metrics span task-specific accuracy measures, computational efficiency indicators, and, increasingly, metrics that attempt to quantify the quality of contextual reasoning itself.

### 5.2.1. Task-Specific Metrics

For object detection tasks, mean Average Precision (mAP) serves as the primary evaluation metric, measuring both the accuracy of object localisation and the confidence of classification. As defined in computational frameworks such as MATLAB's Deep Learning Toolbox, mAP computes the average precision across different recall levels, with detections considered correct when the intersection over union (IoU) between predicted and ground truth bounding boxes exceeds a specified threshold (MathWorks, 2024). The threshold can be varied to assess localisation quality at different levels of strictness, with mAP@0.5 and mAP@0.75 providing insights into coarse and fine localisation performance respectively. For context-aware models, mAP across object categories provides indirect evidence of contextual understanding: accurate detection of small or occluded objects often depends on contextual cues from surrounding scene elements.

## 6. CONCLUSION

This review has systematically examined the landscape of attention-guided deep learning for context-aware image understanding, with particular focus on hierarchical architectures and dynamic feature suppression mechanisms. The evolution from simple channel attention in SENet (Hu et al., 2018) to sophisticated hierarchical Transformers such as Swin Transformer (Liu et al., 2021) demonstrates the rapid progress in enabling networks to selectively emphasise relevant contextual information. Our analysis reveals that while attention mechanisms have matured considerably—encompassing soft and hard attention, channel and spatial operations, and self-attention variants—the integration of dynamic feature suppression remains comparatively underdeveloped. The mathematical unification of attention-suppression blocks presented herein highlights the complementary nature of these mechanisms: attention selects salient features while suppression actively eliminates irrelevant information, yet most existing architectures implement them separately rather than within unified frameworks.

The critical evaluation of benchmarks raises important questions about whether current performance improvements reflect genuine advances in contextual understanding or merely better exploitation of dataset statistics. Liu et al.'s (2025) Context Ambiguity benchmark

highlights the need for evaluation protocols that test models' ability to recognise contextual insufficiency, moving beyond forced-choice accuracy metrics. For dynamic suppression models, interpretability techniques such as those proposed by Ren et al. (2024) offer pathways to verify that suppression targets genuinely irrelevant features rather than discarding useful information. The implementation challenges documented in Python frameworks underscore the gap between research prototypes and production-ready systems, particularly regarding hard attention training stability and dynamic graph optimisation. Future research must therefore pursue unified hierarchical frameworks that jointly optimise attention and suppression, designed with both representational power and deployment efficiency as primary objectives.

### 6.1. Limitations

This review, while comprehensive, acknowledges several limitations that should be considered when interpreting its findings. First, the rapid evolution of attention mechanisms means that recent developments emerging during the review process may not be fully represented, particularly regarding foundation models and large-scale vision-language pre-training that increasingly subsume explicit attention design within broader architectures. Second, the focus on hierarchical attention and dynamic suppression necessarily excludes related paradigms such as neural architecture search and automated attention design, which may offer complementary insights for context-aware understanding. Third, the analysis of benchmark limitations, while critical, does not propose new evaluation protocols but rather synthesises existing critiques, leaving the development of improved benchmarks to future work. Fourth, the discussion of Python implementation challenges reflects current framework capabilities, which continue to evolve rapidly, potentially rendering specific optimisation difficulties transient. Finally, the review's emphasis on architectural patterns may underrepresent the importance of training methodologies, data curation strategies, and regularisation techniques that often prove as crucial as architectural choices for achieving robust context-aware understanding.

## REFERENCES

1. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S. and Zhang, L. (2018) 'Bottom-up and top-down attention for image captioning and visual question answering', IEEE Conference on Computer Vision and Pattern Recognition, pp. 6077-6086.
2. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L. and Parikh, D. (2015) 'VQA: Visual question answering', International Conference on Computer Vision, pp. 2425-2433.

3. Ba, J., Mnih, V. and Kavukcuoglu, K. (2015) 'Multiple object recognition with visual attention', International Conference on Learning Representations.
4. Badrinarayanan, V., Kendall, A. and Cipolla, R. (2017) 'SegNet: A deep convolutional encoder-decoder architecture for image segmentation', IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(12), pp. 2481-2495.
5. Bahdanau, D., Cho, K. and Bengio, Y. (2015) 'Neural machine translation by jointly learning to align and translate', International Conference on Learning Representations.
6. Bengio, Y., Léonard, N. and Courville, A. (2013) 'Estimating or propagating gradients through stochastic neurons for conditional computation', arXiv preprint arXiv:1308.3432.
7. Byeon, W., Breuel, T.M., Raue, F. and Liwicki, M. (2015) 'Scene labeling with LSTM recurrent neural networks', IEEE Conference on Computer Vision and Pattern Recognition, pp. 3547-3555.
8. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K. and Yuille, A.L. (2015) 'Semantic image segmentation with deep convolutional nets and fully connected CRFs', International Conference on Learning Representations.
9. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K. and Yuille, A.L. (2018) 'DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs', IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(4), pp. 834-848.
10. Chen, L.C., Papandreou, G., Schroff, F. and Adam, H. (2017) 'Rethinking atrous convolution for semantic image segmentation', arXiv preprint arXiv:1706.05587.
11. Chen, M., Wu, J., Wang, L. and Zhang, X. (2021) 'Learning to suppress significant activation values for robust image classification', IEEE Transactions on Image Processing, 30, pp. 7892-7905.
12. Child, R., Gray, S., Radford, A. and Sutskever, I. (2019) 'Generating long sequences with sparse transformers', arXiv preprint arXiv:1904.10509.
13. Di Bello, F., Ben Hadj Hassen, S., Astrand, E. and Ben Hamed, S. (2021) 'A unified neurocognitive model of proactive and reactive attentional suppression', Neuroscience & Biobehavioral Reviews, 129, pp. 345-358.
14. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J. and Houtsby, N. (2021) 'An image is worth 16x16 words: Transformers for image recognition at scale', International Conference on Learning Representations.
15. Dutta, A., Mehrab, K.S., Sawhney, M., Neog, A., Khurana, M., Fatemi, S., Pradhan, A., Maruf, M., Lourentzou, I., Daw, A. and Karpatne, A. (2025) 'Open world scene graph generation using vision language models', arXiv preprint arXiv:2506.08189.
16. Farabet, C., Couprie, C., Najman, L. and LeCun, Y. (2013) 'Learning hierarchical features for scene labeling', IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(8), pp. 1915-1929.
17. Figurnov, M., Collins, M.D., Zhu, Y., Zhang, L., Huang, J., Vetrov, D. and Salakhutdinov, R. (2017) 'Spatially adaptive computation time for residual networks', IEEE Conference on Computer Vision and Pattern Recognition, pp. 1039-1048.
18. Gao, X., Zhao, Y., Dudziak, Ł., Mullins, R. and Xu, C.Z. (2019) 'Dynamic channel pruning: Feature boosting and suppression', International Conference on Learning Representations.
19. González-Chávez, O., Ruiz, G., Moctezuma, D. and Ramirez-delReal, T. (2023) 'Are metrics measuring what they should? An evaluation of Image Captioning task metrics', Signal Processing: Image Communication, 117, 107071.
20. Google(2024) 'tf.compat.v1.metrics.mean\_iou', TensorFlow Documentation v2.15.0. Available at: [https://www.tensorflow.org/versions/r2.15/api\\_docs/python/tf/compat/v1/metrics/mean\\_iou](https://www.tensorflow.org/versions/r2.15/api_docs/python/tf/compat/v1/metrics/mean_iou) (Accessed: 19 February 2026).
21. Graham, B., Engelcke, M. and van der Maaten, L. (2018) '3D semantic segmentation with submanifold sparse convolutional networks', IEEE Conference on Computer Vision and Pattern Recognition, pp. 9224-9232.
22. Hassanin, M., Anwar, S., Radwan, I. and Khan, F.S. (2024) 'Visual attention methods in deep learning: A comprehensive survey', ACM Computing Surveys, 56(7), pp. 1-42.
23. He, K., Zhang, X., Ren, S. and Sun, J. (2015) 'Spatial pyramid pooling in deep convolutional networks for visual recognition', IEEE Transactions on Pattern Analysis and Machine Intelligence, 37(9), pp. 1904-1916.
24. He, K., Zhang, X., Ren, S. and Sun, J. (2016) 'Deep residual learning for image recognition', IEEE

- Conference on Computer Vision and Pattern Recognition, pp. 770-778.
25. Hu, J., Shen, L. and Sun, G. (2018) 'Squeeze-and-excitation networks', IEEE Conference on Computer Vision and Pattern Recognition, pp. 7132-7141.
26. Itti, L., Koch, C. and Niebur, E. (1998) 'A model of saliency-based visual attention for rapid scene analysis', IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(11), pp. 1254-1259.
27. Jaderberg, M., Simonyan, K., Zisserman, A. and Kavukcuoglu, K. (2015) 'Spatial transformer networks', Advances in Neural Information Processing Systems, 28, pp. 2017-2025.
28. Jang, E., Gu, S. and Poole, B. (2017) 'Categorical reparameterization with Gumbel-softmax', International Conference on Learning Representations.
29. Jia, X., De Brabandere, B., Tuytelaars, T. and Gool, L.V. (2016) 'Dynamic filter networks', Advances in Neural Information Processing Systems, 29, pp. 667-675.
30. Kim, J.H., Jun, J. and Zhang, B.T. (2018) 'Bilinear attention networks', Advances in Neural Information Processing Systems, 31, pp. 1564-1574.
31. Kitaev, N., Kaiser, Ł. and Levskaya, A. (2020) 'Reformer: The efficient transformer', International Conference on Learning Representations.
32. Krizhevsky, A., Sutskever, I. and Hinton, G.E. (2012) 'ImageNet classification with deep convolutional neural networks', Advances in Neural Information Processing Systems, 25, pp. 1097-1105.
33. Lawson, D., Qureshi, A.S., Chung, W.H. and Bengio, Y. (2018) 'Learning hard attention for visual question answering with discrete optimization', NeurIPS Workshop on Relational Representation Learning.
34. Lazebnik, S., Schmid, C. and Ponce, J. (2006) 'Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories', IEEE Conference on Computer Vision and Pattern Recognition, pp. 2169-2178.
35. Lefaudeux, B., Massa, F., Liskovich, D., Xiong, W., Caggiano, V., Naren, S., Xu, M., Hu, J., Tintore, M., Zhang, S. and LeGendre, C. (2022) 'xformers: A modular and hackable Transformer modelling library', GitHub repository. Available at: <https://github.com/facebookresearch/xformers>.
36. Li, H., Xiong, P., An, J. and Wang, L. (2018) 'Pyramid attention network for semantic segmentation', arXiv preprint arXiv:1805.10180.
37. Li, X., Wang, W., Hu, X. and Yang, J. (2019) 'Selective kernel networks', IEEE Conference on Computer Vision and Pattern Recognition, pp. 510-519.
38. Lin, J., Rao, Y., Lu, J. and Zhou, J. (2017) 'Runtime neural pruning', Advances in Neural Information Processing Systems, 30, pp. 2181-2191.
39. Liu, J., Wang, Y., Zhang, L. and Chen, T. (2025) 'Detecting multimodal situations with insufficient context and abstaining from baseless predictions', arXiv preprint arXiv:2405.11145.
40. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S. and Guo, B. (2021) 'Swin Transformer: Hierarchical vision transformer using shifted windows', International Conference on Computer Vision, pp. 10012-10022.
41. Long, J., Shelhamer, E. and Darrell, T. (2015) 'Fully convolutional networks for semantic segmentation', IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431-3440.
42. Luo, W., Li, Y., Urtasun, R. and Zemel, R. (2016) 'Understanding the effective receptive field in deep convolutional neural networks', Advances in Neural Information Processing Systems, 29, pp. 4898-4906.
43. MathWorks (2024) 'mAPObjectDetectionMetric', MATLAB Deep Learning Toolbox Documentation. Available at: <https://uk.mathworks.com/help/vision/ref/mapobjectdetectionmetric.html> (Accessed: 19 February 2026).
44. Mnih, V., Heess, N., Graves, A. and Kavukcuoglu, K. (2014) 'Recurrent models of visual attention', Advances in Neural Information Processing Systems, 27, pp. 2204-2212.
45. Nature Scientific Reports (2024) 'Table 3: Quantitative comparison of unsupervised methods on Cityscapes, ADE20K, and COCO-Stuff', Nature Scientific Reports.
46. Nie, M., Sun, J., Guoyang, H., Niu, A., Hu, Y., Yan, Q. and Zhu, Y. (2025) 'FSCFNet: Lightweight neural networks via multi-dimensional importance-aware optimization', Neurocomputing, 131823.
47. Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N.Y., Kainz, B., Glocker, B. and Rueckert, D. (2018) 'Attention U-Net: Learning where to look for the pancreas', Medical Imaging with Deep Learning.

48. Oliva, A. and Torralba, A. (2007) 'The role of context in object recognition', Trends in Cognitive Sciences, 11(12), pp. 520-527.
49. OpenBayes (2024) 'Understanding inhibition through maximally tense images', OpenBayes Trends. Available at: <https://trends.openbayes.com/paper/understanding-inhibition-through-maximally> (Accessed: 19 February 2026).
50. Ren, S., He, K., Girshick, R. and Sun, J. (2017) 'Faster R-CNN: Towards real-time object detection with region proposal networks', IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(6), pp. 1137-1149.
51. Ren, Z.-X., Li, Y., Zhang, H. and Wang, J. (2024) 'Understanding inhibition through maximally tense images', arXiv preprint arXiv:2406.08924.
52. Ronneberger, O., Fischer, P. and Brox, T. (2015) 'U-Net: Convolutional networks for biomedical image segmentation', International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 234-241.
53. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C. and Fei-Fei, L. (2015) 'ImageNet large scale visual recognition challenge', International Journal of Computer Vision, 115(3), pp. 211-252.
54. Sigurdson, P. (2024) 'Differences between PyTorch and TensorFlow in AI model development', Coda. Available at: <https://coda.io/@peter-sigurdson/differences-between-pytorch-and-tensorflow-in-ai-model-developme>.
55. Simonyan, K. and Zisserman, A. (2015) 'Very deep convolutional networks for large-scale image recognition', International Conference on Learning Representations.
56. Singapore University of Technology and Design (2025) 'Training indoor and scene-specific semantic segmentation models', IEEE Xplore.
57. Sun, K., Xiao, B., Liu, D. and Wang, J. (2019) 'Deep high-resolution representation learning for human pose estimation', IEEE Conference on Computer Vision and Pattern Recognition, pp. 5693-5703.
58. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A. and Jégou, H. (2021) 'Training data-efficient image transformers&distillationthrough attention', International Conference on Machine Learning, pp. 10347-10357.
59. University of Queensland (2024) 'Comparing ML libraries', HYPPODocumentation. Available at: <https://hpo-uq.gitlab.io/hyppo/architecture/compare.html>.
60. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I. (2017) 'Attention is all you need', Advances in Neural Information Processing Systems, 30, pp. 5998-6008.
61. Vinyals, O., Toshev, A., Bengio, S. and Erhan, D. (2015) 'Show and tell: A neural image caption generator', IEEE Conference on Computer Vision and Pattern Recognition, pp. 3156-3164.
62. Visin, F., Ciccone, M., Romero, A., Kastner, K., Cho, K., Bengio, Y., Matteucci, M. and Courville, A. (2016) 'Reseg: A recurrent neural network-based model for semantic segmentation', CVPR Workshop on Deep Learning for Semantic Segmentation.
63. Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., Liu, W. and Xiao, B. (2020) 'Deep high-resolution representation learning for visual recognition', IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(10), pp. 3349-3364.
64. Wang, X., Girshick, R., Gupta, A. and He, K. (2018) 'Non-local neural networks', IEEE Conference on Computer Vision and Pattern Recognition, pp. 7794-7803.
65. Wightman, R. (2025) 'timm: PyTorch image models', GitHub repository. Available at: <https://github.com/huggingface/pytorch-image-models>.
66. Woo, S., Park, J., Lee, J.Y. and Kweon, I.S. (2018) 'CBAM: Convolutional block attention module', European Conference on Computer Vision, pp. 3-19.
67. Wu, D., Li, Z. and Mitra, T. (2025) 'Inkstream: Instantaneous GNN inference on dynamic graphs via incremental update', IEEE International Parallel and Distributed Processing Symposium.
68. Wu, H., Xiao, B., Codella, N., Liu, M., Dai, X., Yuan, L. and Zhang, L. (2021) 'CvT: Introducing convolutions to vision transformers', International Conference on Computer Vision, pp. 22-31.
69. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M. and Luo, P. (2021) 'SegFormer: Simple and efficient design for semantic segmentation with transformers', Advances in Neural Information Processing Systems, 34, pp. 12077-12090.
70. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R. and Bengio, Y. (2015)

'Show, attend and tell: Neural image caption generation with visual attention', International Conference on Machine Learning, pp. 2048-2057.

71. Yang, B., Bender, G., Le, Q.V. and Ngiam, J. (2019) 'CondConv: Conditionally parameterized convolutions for efficient inference', Advances in Neural Information Processing Systems, 32, pp. 1307-1318.
72. Yang, Z., He, X., Gao, J., Deng, L. and Smola, A. (2016) 'Stacked attention networks for image question answering', IEEE Conference on Computer Vision and Pattern Recognition, pp. 21-29.
73. You, Z., Wang, J., Kong, L., He, B. and Wu, Z. (2025) 'Pix2Cap-COCO: Advancing visual comprehension via pixel-level captioning', arXiv preprint arXiv:2501.13893.
74. Yuan, Y., Chen, X. and Wang, J. (2020) 'Object-contextual representations for semantic segmentation', European Conference on Computer Vision, pp. 173-190.
75. Zhang, H., Niu, Y. and Chang, S.F. (2019) 'Hierarchical attention networks for image captioning', AAAI Conference on Artificial Intelligence, 33, pp. 9253-9260.
76. Zhang, H., Li, F., Liu, S., Zhang, L., Su, H., Zhu, J., Ni, L.M. and Shum, H.Y. (2025) 'Optimising vision transformer performance on limited datasets: A multi-gradient approach', IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops.
77. Zhao, H., Shi, J., Qi, X., Wang, X. and Jia, J. (2017) 'Pyramid scene parsing network', IEEE Conference on Computer Vision and Pattern Recognition, pp. 2881-2890.
78. Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., Fu, Y., Feng, J., Xiang, T., Torr, P.H. and Zhang, L. (2021) 'Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers', IEEE Conference on Computer Vision and Pattern Recognition, pp. 6881-6890.