

Attentive Convolution: Unifying the Expressivity of Self-Attention with Convolutional Efficiency

Hao Yu, Haoyu Chen, Yan Jiang, Wei Peng, Zhaodong Sun, Samuel Kaski, Guoying Zhao, *Fellow, IEEE*

Abstract—Self-attention (SA) has become the cornerstone of modern vision backbones for its powerful expressivity over traditional Convolutions (Conv). However, its quadratic complexity remains a critical bottleneck for practical applications. Given that Conv offers linear complexity and strong visual priors, continuing efforts have been made to promote the renaissance of Conv. However, a persistent performance chasm remains, highlighting that these modernizations have not yet captured the intrinsic expressivity that defines SA. In this paper, we re-examine the design of the CNNs, directed by a key question: what principles give SA its edge over Conv? As a result, we reveal two fundamental insights that challenge the long-standing design intuitions in prior research (e.g., Receptive field). The two findings are: (1) *Adaptive routing*: SA dynamically regulates positional information flow according to semantic content, whereas Conv employs static kernels uniformly across all positions. (2) *Lateral inhibition*: SA induces score competition among token weighting, effectively suppressing redundancy and sharpening representations, whereas Conv filters lack such inhibitory dynamics and exhibit considerable redundancy. Based on this, we propose *Attentive Convolution* (ATConv), a principled reformulation of the convolutional operator that intrinsically injects these principles. Interestingly, with only 3×3 kernels, ATConv consistently outperforms various SA mechanisms in fundamental vision tasks. Building on ATConv, we introduce AttNet, a CNN family that can attain 84.4% ImageNet-1K Top-1 accuracy with only 27M parameters. In diffusion-based image generation, replacing all SA with the proposed 3×3 ATConv in SiT-XL/2 reduces ImageNet FID by 0.15 in 400k steps with faster sampling. Code is available at: github.com/price112/Attentive-Convolution.

Index Terms—Adaptive routing, lateral inhibition, convolution, self-attention, vision transformer.

I. INTRODUCTION

Convolutional neural networks (CNNs) [1]–[6] have long dominated computer vision, achieving remarkable success across diverse tasks owing to their inherent visual inductive biases and computational efficiency on high-dimensional inputs. Recently, Vision Transformers (ViTs) [7]–[12] have emerged as a strong competitive alternative, leveraging the

self-attention mechanism [13] to model global dependencies through content-based query–key interactions. Unlike convolutions, which encode fixed local patterns with limited receptive fields, self-attention enables flexible long-range modeling. However, vanilla self-attention suffers from quadratic computational complexity with respect to the input resolution, rendering it inefficient for visual data characterized by high dimensionality and substantial encoding redundancy. Despite that the ViT series [7] leverages large patch sizes (e.g., 16×16 or 32×32) to reduce the sequence length of input images, its computational complexity is still much higher than that of traditional CNNs. Furthermore, self-attention’s position-agnostic design treats all spatial locations equally during pixel dependency modeling, requiring extensive training resources to learn fundamental visual priors such as locality and object continuity from scratch.

To address these limitations, modern ViTs have undergone substantial architectural evolution. An important revolution was the adoption of CNN-like pyramid designs [9], [12], which progressively downsample spatial dimensions, restricting attention computation to manageable resolutions. Based on this, recent state-of-the-art ViTs [14]–[18] further hybridize attention with depth-wise convolutions [19], [20], creating models that strategically leverage convolution’s efficiency for local processing while preserving attention’s capacity for global modeling. Notably, the persistence of convolution reveals a crucial insight: its visual efficiency and inherent visual inductive biases remain indispensable for vision systems.

Recognizing this complementarity, researchers have sought to augment CNNs with ViT-inspired principles. The most common idea is to expand the receptive field that approximates the global modeling of attention. ConvNeXt [21], [22] modernizes ConvNets with ViT design elements and adopts larger kernels instead of the traditional 3×3 ones, while reparameterization-based architectures [23], [24] further extend kernel sizes up to 31×31 to capture long-range dependencies. Although these strategies yield notable improvements over classical CNNs such as ResNet [1], a substantial gap from modern ViTs persists. For example, with similar parameters, TransNeXt-Tiny [17] wins ConvNeXt-Tiny [21] by 2.0% Top-1 accuracy in ImageNet-1K classification. This disparity highlights that simply augmenting CNN from the receptive field perspective is insufficient to capture the fundamental advantages of ViTs.

In this paper, we dive deeper and present two new perspectives to understand the underlying visual expressivity of self-attention over Conv-based operators: the *adaptive routing* and *lateral inhibition* properties. Through a unified weighted aggregation framework, we demonstrate that

This work was supported by the Research Council of Finland Academy Professor project EmotionAI (grants 336116, 345122, 359854), the University of Oulu & Research Council of Finland Profi 7 (grant 352788), and Academy of Finland Flagship program: the Finnish Center for Artificial Intelligence FCAI. (Corresponding Author: Guoying Zhao)

H. Yu, H. Chen, Y. Jiang, G. Zhao are with Center for Machine Vision and Signal Analysis, University of Oulu, Finland (e-mail: hao.2.yu@oulu.fi; chen.haoyu@oulu.fi; yan.jiang@oulu.fi; guoying.zhao@oulu.fi).

W. Peng is with the Department of Psychiatry and Behavioral Sciences, Stanford University, USA. (e-mail: wepeng@stanford.edu).

Z. Sun is with the School of Computer Science, Nanjing University of Information Science and Technology, China (e-mail: zhaodong.sun@nuist.edu.cn).

S. Kaski is with the Department of Computer Science, Aalto University, Finland. (e-mail: samuel.kaski@aalto.fi).

S. Kaski and G. Zhao are with the ELLIS Institute Finland (first affiliation).

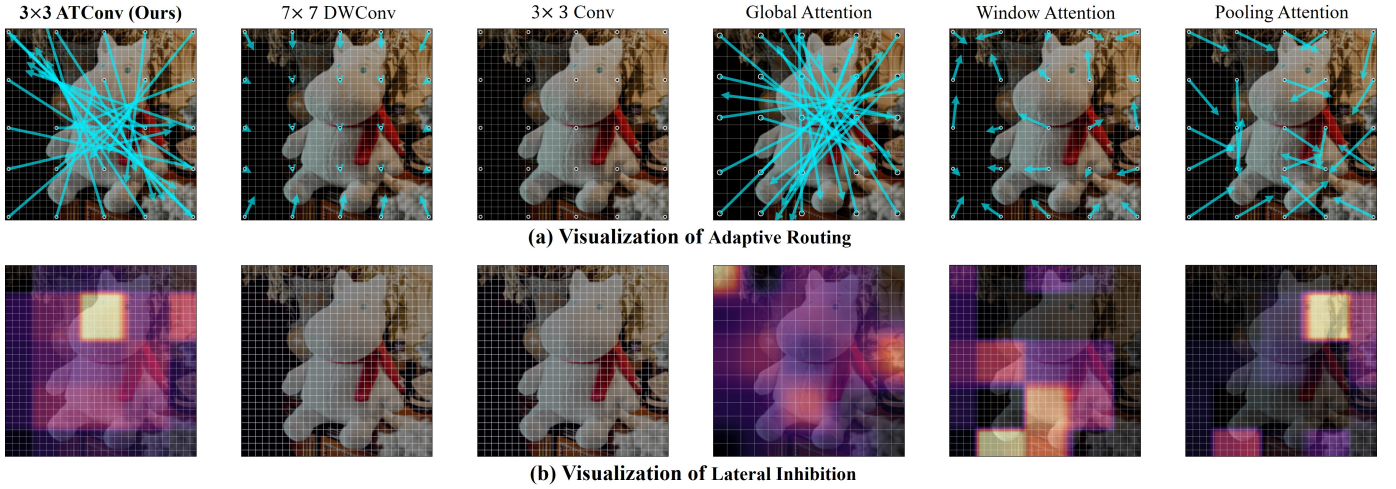


Fig. 1. Visualization of adaptive routing and lateral inhibition on 3×3 ATConv, 7×7 DWConv [21], 3×3 Conv [1], vanilla global attention [7], window [9] and pooling [12] attention. We analyze operator-intrinsic behavior by extracting influence maps $G(h, w) = \sum_c |\partial y_{c^*, h^*, w^*} / \partial x_{c, h, w}|$. Adaptive routing is visualized via distance-weighted centroids of high-influence regions, with FAR at radial threshold r_0 measuring long-range preference. Diverse arrows indicate different inputs produce different aggregation patterns, while uniform arrows reveal fixed routing. Lateral inhibition is quantified by the off-center suppression response to positive perturbations at anchors, where brighter regions indicate stronger surround suppression. ATConv with a compact 3×3 kernel exhibits pronounced adaptive routing and lateral inhibition effects similar to the global self-attention. See our depository for detailed visualization code.

vanilla convolution is a static weight aggregation operator with an identity representation basis, which leverages the fixed pixel aggregation rule across all spatial positions. In contrast, self-attention implements *adaptive routing* via query-key interactions with learnable basis space transformations. This adaptivity enables attention modeling to have selective information transfer based on semantic relevance rather than fixed rules. More importantly, the softmax normalization over attention scores induces crucial *lateral inhibition* dynamics [25] within attention calculation, a canonical mechanism of biological vision where neurons compete for visual selection. In the primary visual cortex (V1) [26], neurons exhibit similar center-surround antagonism, with neighboring neurons competing to represent different features. This competitive dynamic prevents redundant encoding and sharpens feature selectivity—precisely the characteristics lacking in standard convolution. In self-attention mechanisms, the softmax on attention score creates competitive dynamics which implements mutual suppression: increasing the attention weight to one position necessarily decreases others. This sharpens the representations and avoids giving flat responses (e.g., low-rank prediction). The resulting competitive pressure amplifies discriminative connections while suppressing spurious or noisy patterns, leading to more selective and robust representations. In contrast, convolution’s static and independent kernels lack such inhibitory dynamics, fundamentally constraining their expressivity and leading to substantial representation redundancy.

To validate our analysis, we propose *Attentive Convolution* (ATConv), a principled revolution of convolution that embeds the adaptive routing and lateral inhibition principles distilled from self-attention. Unlike vanilla convolutions that employ static, input-agnostic kernels, ATConv adaptively derives its kernels from the input through a context-to-kernel translation mechanism. The key insight is that although convolution operates locally, its kernels can be contextually conditioned to encode global dependencies. In this way, global semantic

relevance is translated into input-dependent kernels, implicitly establishing adaptive routing correspondences through local operations, without incurring the quadratic cost of global pairwise calculation. Beyond routing, we further inject the differential kernel modulation (DKM) into the operation logic of ATConv, which is designed to introduce the lateral inhibition dynamic tailored for the convolutional framework. Inspired by classical descriptors such as LBP [27], [28] and DoG [29], the DKM mechanism performs difference-oriented kernel modulation during convolutional operation, dynamically enhancing feature sharpness while suppressing redundant channel responses.

Fig. 1 empirically validates our findings. In part (a), routing arrows depict how each position aggregates information: uniform arrows denote fixed routing, while diverse arrows indicate adaptive routing for different pixel positions. Traditional Conv operators (3×3 Conv and 7×7 DWConv) show a fixed local routing. Even with a larger kernel, 7×7 DWConv yields identical aggregation patterns across positions, underscoring that kernel enlargement cannot induce adaptivity. In contrast, global self-attention produces diverse routing across spatial regions, while pooling and window attention trade routing capacity for efficiency due to local attention constraints. In contrast, ATConv achieves adaptive routing comparable to global self-attention using only compact 3×3 kernels. Its diverse arrow patterns confirm that the context-to-kernel translation encodes global correspondences into local kernels, enabling convolutional traversal to realize global adaptive routing with even a compact 3×3 kernel size.

Part (b) visualizes the lateral inhibition effect. Standard Conv and DWConv show no inhibitory behavior, highlighting the absence of neuron suppression in traditional convolutional operators. In contrast, ATConv demonstrates competitive dynamics akin to the three compared attention mechanisms, with strong gradient-score competition that sharpens representations and suppresses redundancy.

Building on ATConv, we introduce the Attentive Convolutional Network (AttNet), a purely convolutional architecture that attains state-of-the-art performance while dispensing with self-attention. Moreover, we show that ATConv can serve as a drop-in replacement for self-attention, consistently improving accuracy and efficiency across diverse vision tasks and backbones. Our main contributions are as follows.

- We identify *adaptive routing* and *lateral inhibition* as the key mechanisms behind the superior expressivity of self-attention, and provide both theoretical and empirical evidence that these properties govern representational expressivity.

- We propose *Attentive Convolution* (ATConv), which embeds the adaptive routing and lateral inhibition principles into the convolutional framework, delivering attention-level expressivity with convolutional efficiency.

- We evaluate ATConv across a wide range of discrimination and generation tasks, revealing its consistent advantages over leading attention mechanisms. These results solidify ATConv's position as a new foundational operator, poised to drive the development of next-generation visual models.

II. RELATED WORK

A. Vision Transformer

In the field of natural language processing, Transformer [13] leverages self-attention to model global dependencies between tokens. Vision Transformer (ViT) [7] pioneered the adaptation of this mechanism to computer vision, demonstrating that pure attention-based architectures can achieve competitive performance on image recognition tasks. However, the quadratic complexity of self-attention and the lack of visual inductive biases bring significant computational overhead in visual tasks. The adaptation of self-attention in visual data has motivated extensive research along the following two primary directions.

Computational efficiency. The $O(N^2)$ complexity of self-attention becomes prohibitive for high-resolution inputs like natural images. Pooling-based methods (e.g., PVT [12], PVTv2 [16], P2T [18]) reduce computation through spatial downsampling. Window-based methods restrict attention scope, like Swin Transformer [9] employs shifted windows for linear complexity. Afterwards, more advanced window partition techniques brought further performance improvements, e.g., Cswin [11] uses cross-shaped windows for efficient global modeling, and MaxViT [30] leverages the block-grid interlaced window to promote information flow of local attention. Linear attention variants fundamentally alter the computational paradigm by replacing softmax with kernel functions, exploiting associative properties to achieve $O(N)$ complexity. XCiT [10] computes cross-covariance between feature channels rather than token-wise attention. InLine attention [31] introduces injectivity constraints to preserve discriminative power without using softmax, and CosFormer [32] incorporates cosine-based reweighting for improved stability.

Inductive biases. Self-attention lacks the visual priors inherent to convolutions, necessitating explicit incorporation via positional encodings or Conv integration. Positional encodings provide crucial spatial awareness through various formulations (relative [33], 2D RoPE [34], CPB [35], LePE [36]). More

effective approaches directly combine convolutions: CoAtNet [37] systematically integrates depth-wise convolution with self-attention, while CvT [38] introduces convolutional token embedding and projection layers. InLine [31] attention systematically explains the importance of convolutional locality in attention modeling. To further improve efficiency, FastViT [14] builds a more progressive CNN-ViT hybrid architecture only by using attention in the last stage. Biologically-inspired architectures leverage human vision principles, e.g., Focal Transformer implements coarse-to-fine attention across resolutions, while TransNeXt [17] aggregates multiscale local features through pixel-focused attention.

While extensive efforts have focused on augmenting self-attention with convolutional biases, the dual direction, enhancing convolutions with attention's key advantages, remains underexplored. Since convolution inherently provides visual efficiency, augmenting it with attention's positive principles present a promising direction. This paper shows that transferring self-attention's positive principles to Conv can significantly improve performance.

B. Convolutional Neural Networks

The convolutional operator [39] lies at the heart of modern visual recognition models. By sliding a shared kernel across local neighborhoods, it imposes several strong *visual inductive biases*, including locality, translation equivariance, weight sharing, and hierarchical feature composition. These properties align well with the statistics of natural images. Built on the convolutional operator, Convolutional Neural Networks (CNNs) [1]–[3], [5], [6], [20], [23] have therefore dominated the field of computer vision for decades, evolving from early models such as LeNet [39] to modern architectures such as ResNet [1], DenseNet [5], and EfficientNet [40]. However, with the advent of self-attention in vision tasks, CNNs have been rapidly supplanted by Vision Transformers (ViTs), which exhibit superior performance through the incorporation of self-attention mechanisms with long-range adaptive routing and competitive score modeling capacities.

Representative efforts to narrow down the gap between CNNs and ViTs include ConvNeXt [21], which adopts architectural paradigms from ViTs and enlarges convolutional kernels from 3×3 to 7×7 in order to emulate the long-range modeling capacity of self-attention. Although ConvNeXt shows promising improvements, a substantial performance gap remains compared to the leading ViTs. Subsequent CNNs have similarly emphasized architectural modifications to mimic ViT characteristics, such as employing large kernels (e.g., RepLkNet [24], InceptionNeXt [41]), dilated convolutions [42]. However, these efforts primarily address receptive-field limitations while overlooking a more fundamental issue: the gap stems not merely from spatial coverage, but from the intrinsic differences in visual modeling between convolution and self-attention. This paper steps further by analyzing the key attributes of self-attention in visual modeling and reformatting the Conv operator accordingly, aiming to further mitigate the performance gap between CNNs and ViTs.

C. Dynamic Convolutional Architectures

Early explorations into content-adaptive convolutions aimed to make kernels input-dependent over static schemes. Cond-Conv [43] and DynamicConv [44] learn mixtures of base kernels, while WeightNet [45] and ODConv [46] use hyper-networks to generate dynamic weights. Involution [47] designs spatially-specific yet channel-agnostic filters. SENet [48] and CBAM [49] recalibrate features with attention modules. While these approaches outperform vanilla convolutions, they typically incur high computational/parameter costs and lack the performance scalability needed to serve as fundamental backbone operators. Consequently, most dynamic convolutional designs remain as auxiliary modules rather than being core components of modern architectures.

A deeper limitation of these methods is the absence of *lateral inhibition* [25], the competitive dynamic central to self-attention. Existing dynamic convolutions typically modulate features independently via gates or additive combinations, without enforcing inter-feature competition. This leads to diffuse responses unable to suppress noise or irrelevant signals, explaining why even complex dynamic operators still lag far behind self-attention. ATConv addresses these gaps by incorporating *adaptive routing* to capture content-dependent adaptation and *lateral inhibition* to introduce competitive dynamics. This synergy produces sharper, more discriminative representations reminiscent of self-attention, yet preserves the efficiency and structural simplicity of convolution. Importantly, ATConv's verified generality, efficiency, and scalability establish it as a strong candidate for a foundational operator similar to the self-attention, transcending the limitations of prior dynamic-based approaches.

III. METHODOLOGY

A. Preliminaries

To analyze the intrinsic difference between self-attention and Conv, we first establish a framework where both operators perform weighted aggregation over a signal manifold. Let $\mathbf{X} \in \mathbb{R}^{B \times C \times H \times W}$ denote the input signal with spatial dimensions $H \times W$ and feature (channel) dimension C . For analysis, we also use the flattened view $\mathbf{X} \in \mathbb{R}^{B \times N \times C}$ with $N = H \times W$. **Definition 1: Generalized Aggregation Operator.** Given the input signal $\mathbf{X}$, the output at position i is:

$$\mathbf{y}_i = \sum_{j \in \Omega_i} \alpha_{ij} \cdot \mathcal{T}(\mathbf{X}_j), \quad (1)$$

where Ω_i denotes the aggregation domain, α_{ij} represents the aggregation weight from position j to position i , and $\mathcal{T} : \mathbb{R}^C \rightarrow \mathbb{R}^{C'}$ is the basis transform at position j . The crucial distinction between self-attention and Conv lies in how α_{ij} and $\mathcal{T}_j$ are determined. In the following, we analyze Conv in its *depthwise* form, which uses a scalar weight per spatial offset for each channel and is more comparable to self-attention's content-weighted summation.

B. The Fundamental Distinction in Signal Aggregation

Adaptive Routing vs. Static Aggregation. We identify adaptive routing as the fundamental distinction between self-attention and convolution. At its core, *routing* refers to how

information flows from input positions to output during pixel aggregation, encompassing both which positions contribute (α_{ij}) and what representations they provide ($\mathcal{T}_j$). Our insight is that achieving adaptive routing requires both components to adapt dynamically based on input content.

Conv employs *static routing*, where aggregation are fixed regardless of input:

$$\mathbf{y}_i^{\text{Conv}} = \sum_{j \in \mathcal{N}_K(i)} w_{p(i,j)} \mathbf{x}_j, \quad (2)$$

here, $w_{p(i,j)}$ denotes the weight of the kernel in relative position $p(i,j) = j - i$, and $\mathcal{N}_K(i)$ represents the local neighborhood $K \times K$. These weights remain input-independent ($\partial w_{p(i,j)} / \partial \mathbf{x} = \mathbf{0}$), and the basis features undergo no transformation mapping before aggregation ($\mathcal{T}_j^{\text{Conv}}(\mathbf{X}) = \mathbf{x}_j$). Consequently, convolution applies identical aggregation patterns universally: a 3×3 edge detector employs the same weights whether processing edges, textures, or uniform regions, fundamentally limiting its adaptability.

Self-attention, instead, achieves *adaptive routing* through two synergistic mechanisms:

$$\begin{aligned} \mathbf{y}_i^{\text{SA}} &= \sum_{j=1}^N \alpha_{ij}^{\text{SA}}(\mathbf{X}) \cdot \mathbf{v}_j, \\ \alpha_{ij}^{\text{SA}} &= \frac{\exp(s_{ij}/\tau)}{\sum_{m=1}^N \exp(s_{im}/\tau)}, \end{aligned} \quad (3)$$

where $s_{ij} = \langle \mathbf{q}_i, \mathbf{k}_j \rangle$ quantifies query-key affinity. The routing weights α_{ij}^{SA} adapt dynamically based on content similarity, while the value $\mathbf{v}_j = \mathbf{W}_v \mathbf{x}_j$ (the $\mathcal{T}$ in Eq. 1) projects basis features into an optimized representation space. This dual adaptation makes weights determine *where* to aggregate based on semantic relevance, while value projection determines *what* representations to aggregate for better information routing.

To further establish why both components are necessary, we analyze the sensitivity of each operator to input perturbations with the Kronecker delta δ_{jn} :

$$\begin{aligned} \frac{\partial \mathbf{y}_i^{\text{Conv}}}{\partial \mathbf{x}_n} &= \begin{cases} w_{p(i,n)} \cdot \mathbf{I}_{C'} & \text{if } n \in \mathcal{N}_K(i) \\ \mathbf{0}_{C'} & \text{otherwise,} \end{cases} \\ \frac{\partial \mathbf{y}_i^{\text{SA}}}{\partial \mathbf{x}_n} &= \sum_{j=1}^N \left[\frac{\partial \alpha_{ij}^{\text{SA}}}{\partial \mathbf{x}_n} \cdot \mathbf{v}_j + \alpha_{ij}^{\text{SA}} \cdot \delta_{jn} \cdot \mathbf{W}_v \right]. \end{aligned} \quad (4)$$

This sensitivity analysis reveals a fundamental computational hierarchy. The static weights of Conv yield gradients invariant to input content, constraining it to uniform spatial processing on untransformed features. In contrast, self-attention's sensitivity decomposes into two adaptive components that enable content-aware computation. The first term, $(\partial \alpha_{ij}^{\text{SA}} / \partial \mathbf{x}_n) \mathbf{v}_j$, facilitates adaptive routing by adjusting aggregation weights based on input characteristics. The second term, $\alpha_{ij}^{\text{SA}} \cdot \mathbf{W}_v$, represents learned value transformations that project features into task-optimized subspaces. This transformation is critical: without it ($\mathbf{W}_v = \mathbf{I}$), the term degenerates to $\alpha_{in}^{\text{SA}} \cdot \mathbf{I}_{C'}$, restricting aggregation to the original feature space where semantic concepts may be poorly separated. Value projections enable discovering task-specific manifolds

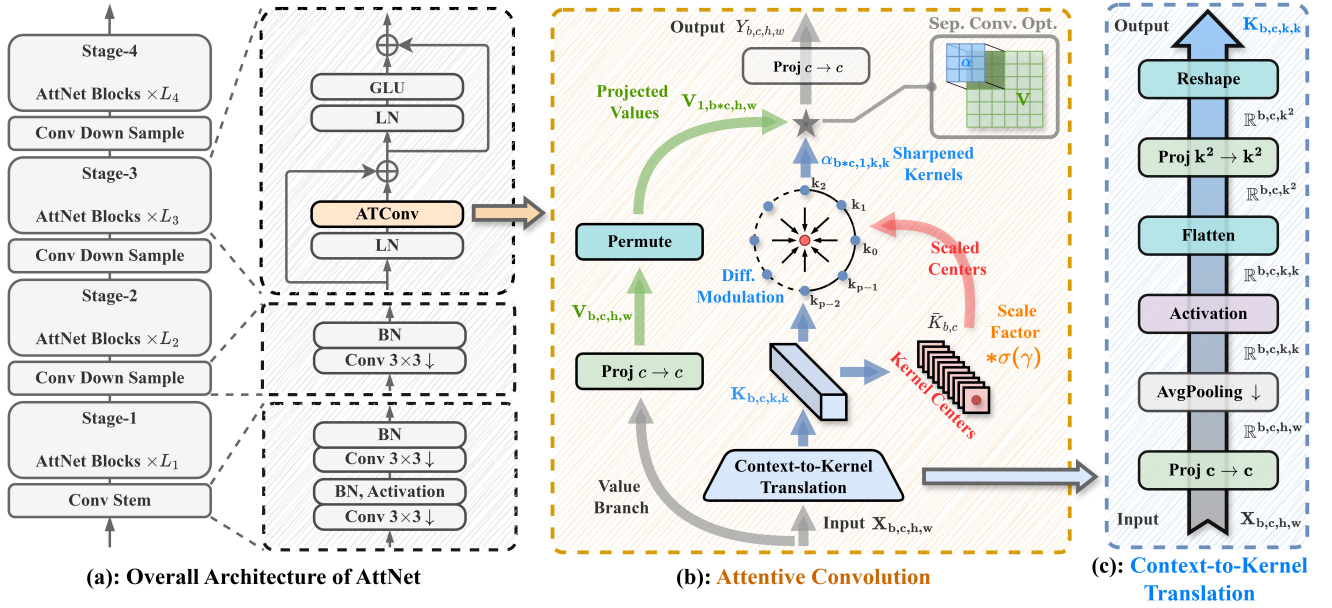


Fig. 2. (a) Overall architecture of AttNet, where we use ATConv as token-mixer (spatial-operator) and GLU as channel-mixer; (b) Architecture of Attentive Convolution (ATConv); (c) Architecture of kernel generator which builds the initial kernels for ATConv based on input contents.

where similar concepts cluster and dissimilar ones separate, facilitating more effective aggregation.

This analysis establishes a clear progression in routing capabilities. Vanilla Conv implements neither adaptive components, operating with fixed kernels on untransformed features. Existing dynamic Convs advance halfway by generating content-dependent weights, yet still aggregate within the original feature space. Self-attention alone achieves complete adaptive routing through both (i) content-dependent weights that determine where to gather information and (ii) learnable feature space transformations that optimize what representations to aggregate. This progression from static to partial to complete adaptation explains the persistent performance gap between convolution and attention in visual tasks.

Lateral Inhibition: Competitive Dynamics in Aggregation. Beyond adaptive routing, we analyze the lateral inhibition between aggregation weights. Convolution weights operate independently without mutual influence:

$$w_{p(i,j)} = \text{constant}, \quad \sum_{j \in N_K(i)} w_{p(i,j)} \neq 1 \text{ (generally)}. \quad (5)$$

Each kernel weight functions in isolation, modifying one weight does not affect others. This independence means all positions contribute according to fixed weights regardless of their relative importance, potentially aggregating both signal and noise with equal emphasis for different inputs. Furthermore, without competitive dynamics to enforce specialization, independent kernels often converge to similar features, producing redundant filters. Multiple channels may learn nearly identical edge detectors or texture filters, resulting in a low rank transformation where hundreds of parameters encode only dozens of unique patterns. This representational redundancy, exacerbated by the absence of lateral inhibition, yields flat, diffuse responses that waste the network's capacity.

Self-attention implements competitive dynamics through softmax normalization, creating mutual inhibition between aggregation weights:

$$\frac{\partial \alpha_{ij}^{\text{SA}}}{\partial s_{ik}} = \begin{cases} \alpha_{ij}^{\text{SA}}(1 - \alpha_{ij}^{\text{SA}})/\tau & \text{if } j = k \\ -\alpha_{ij}^{\text{SA}}\alpha_{ik}^{\text{SA}}/\tau & \text{if } j \neq k \end{cases} \quad (6)$$

The negative off-diagonal terms $-\alpha_{ij}^{\text{SA}}\alpha_{ik}^{\text{SA}}$ implement lateral inhibition: increasing the affinity score s_{ik} for one position necessarily decreases weights α_{ij} for all other positions $j \neq k$. This creates a dynamic where positions compete for aggregation bandwidth, forcing each attention score to specialize on distinct patterns rather than redundantly encoding similar features. Unlike convolution's independent weights that often converge to similar filters, this competitive pressure ensures diverse representations across heads and positions. The strongest semantic connections amplify while weaker ones suppress, yielding sharp, non-redundant feature maps with high effective rank. This lateral inhibition mechanism thus enhances both representational quality and robustness, as the network learns complementary features that capture different aspects of the input rather than wasting capacity on duplicate patterns.

Summary. Adaptive routing and lateral inhibition fundamentally distinguish self-attention from convolution, extending beyond differences in receptive field size. These mechanisms enable self-attention from (1) adaptive information flow that responds to semantic content rather than following fixed spatial patterns, and (2) competitive selection that amplifies task-relevant signals and reduces feature redundancy, while suppressing noise. These properties are hypothesized to contribute to self-attention's strong empirical performance on complex vision tasks requiring selective aggregation and global context integration, motivating the development of convolution operators that incorporate these routing capabilities while preserving computational efficiency.

C. Attentive Convolution: From Theory to Design

Our analysis shows that the strength of self-attention arises from two principles absent in convolution: *adaptive routing*, which enables content-aware aggregation, and *lateral inhibition*, which enables competitive dynamics. The key challenge is to translate these abstract principles into a convolutional framework. Guided by Eq. 4 and Eq. 6, ATConv instantiates these principles through three principled revolutions based on the vanilla convolution framework: (i) a *context-to-kernel translation* mechanism generating routing weights (kernel) that encode the global semantic understanding into local processing rules; (ii) a learnable *value projection* for basis adaptation; and (iii) a *differential kernel modulation* injecting lateral inhibition between kernel entries via difference-oriented modulation.

Adaptive Routing with Context-to-Kernel Translation. To enable adaptive routing within the convolutional framework, ATConv introduces a *Context-to-Kernel Translation* (C2K) mechanism that technically departs from conventional kernel designs. C2K functions as a semantic compiler that bridges global scene understanding with local processing rules. This is reached by encoding the complete $H \times W$ spatial context into compact semantic representations and translating these into tailored filtering operations. Given an input tensor $\mathbf{X} \in \mathbb{R}^{B \times C \times H \times W}$, the C2K executes the following steps:

$$\begin{aligned} \mathbf{F} &= \text{Conv}_{1 \times 1}(\mathbf{X}), & \mathbf{F} &\in \mathbb{R}^{B \times C \times H \times W}, \\ \mathbf{Z} &= \text{AdaAvgPool}_{K \times K}(\mathbf{F}), & \mathbf{Z} &\in \mathbb{R}^{B \times C \times K \times K}, \\ \hat{\mathbf{K}} &= \mathbf{W}_{\text{gen}} \cdot \text{Vec}(\phi(\mathbf{Z})), & \hat{\mathbf{K}} &\in \mathbb{R}^{B \times C \times K^2}, \\ \mathbf{K} &= \text{Reshape}(\hat{\mathbf{K}}), & \mathbf{K} &\in \mathbb{R}^{B \times C \times K \times K}. \end{aligned} \quad (7)$$

The above pipeline proceeds through four key steps. First, a pointwise convolution $\text{Conv}_{1 \times 1}$ projects the input features into a routing-aware latent space, encoding global contextual information for kernel synthesis. Second, adaptive average pooling $\text{AdaAvgPool}_{K \times K}$ compresses the full $H \times W$ spatial resolution into a compact $K \times K$ representation, where each position corresponds to a kernel coefficient location. Third, after vectorizing ($\mathbb{R}^{K \times K} \rightarrow \mathbb{R}^{K^2}$) this pooled representation via $\text{Vec}(\cdot)$, we apply a channel-shared linear transformation $\mathbf{W}_{\text{gen}} \in \mathbb{R}^{K^2 \times K^2}$ with nonlinear activation ϕ (e.g., GELU) to translate semantic codes into kernel entities. Finally, the result vector is reshaped to recover the $K \times K$ spatial structure.

This design embodies the core principle of *Context-to-Kernel Translation*: global scene knowledge is first distilled into a compact semantic encoding, then re-expressed as spatially-adaptive filtering operations. When convolution occurs at position (h, w) , the generated kernel $\mathbf{K}_{b,c,;}$ aggregates local neighborhoods using weights informed by the entire spatial context. As the convolution window traverses the feature map, each location receives a uniquely tailored kernel that captures how its local structure relates to the global scene, effectively establishing an implicit routing network.

Our empirical results in Fig. 1 and Tab. IX-(b) reveal that C2K enables even compact 3×3 kernels to capture long-range dependencies at the 224px scale. This finding highlights a crucial insight: context-aware kernel generation through C2K provides a more effective and efficient pathway to global modeling, rather than mechanically increasing kernel sizes.

Lateral Inhibition via Differential Kernel Modulation. Our analysis reveals that self-attention derives its representational power not only from adaptive routing, but also crucially from the inherent *lateral inhibition* dynamics (Eq. 6). Thus, we introduce *Differential Kernel Modulation* (DKM), a mechanism that explicitly incorporates lateral inhibition into ATConv. The DKM modulates the final ATConv kernels α^{ATConv} as follows:

$$\begin{aligned} \alpha_{b,c,u,v}^{\text{ATConv}} &= \mathbf{K}_{b,c,u,v} - \lambda_c \bar{\mathbf{K}}_{b,c}, \\ \text{where } \bar{\mathbf{K}}_{b,c} &= \frac{1}{K^2} \sum_{u,v} \mathbf{K}_{b,c,u,v}, \\ \lambda_c &= \sigma(\gamma_c) \in (0, 1). \end{aligned} \quad (8)$$

Here, $\gamma \in \mathbb{R}^C$ is a learnable vector that, through the sigmoid function $\sigma(\cdot)$, produces channel-specific inhibition coefficients $\lambda_c \in (0, 1)$. $\bar{\mathbf{K}}_{b,c}$ is the spatial mean of the kernel weight. By subtracting $\lambda_c \bar{\mathbf{K}}_{b,c}$ from each kernel position, DKM transforms absolute values into differential signals centered around their spatial average adaptively. This mechanism instantiates the principle of *center-surround antagonism* fundamental to biological vision. In the primary visual cortex, neurons compute responses as central excitation minus weighted surround inhibition, suppressing redundant patterns while preserving salient features. DKM translates this neurobiological principle computationally, with learnable λ_c that enables adaptive inhibition profiles for each individual kernel: strong suppression ($\lambda_c \rightarrow 1$) for edge detection, moderate for texture discrimination, and weak ($\lambda_c \rightarrow 0$) for smooth gradient preservation.

A distinguishing characteristic of DKM is its *kernel-wise heterogeneity*. Each kernel maintains independent reference means $\bar{\mathbf{K}}_{b,c}$ and inhibition coefficients λ_c , establishing diverse competitive dynamics across the feature space. This prevents collapse and promotes functional specialization: high-frequency kernels naturally evolve stronger inhibition to sharpen boundaries, while semantic kernels maintain weaker inhibition for holistic pattern capture. Such differentiation enhances the effective rank of representations by ensuring non-redundant channel responses, while also stabilizing training by preventing degenerate flat activations that can plague deeper layers. Through this biologically-inspired mechanism, DKM transforms standard convolution into a contrast-sensitive operation that balances local detail with global context.

This mechanism is further clarified by the Jacobian structure of DKM with respect to $\mathbf{K}$:

$$\frac{\partial \alpha_{b,c,u,v}^{\text{ATConv}}}{\partial \mathbf{K}_{b,c,u',v'}} = \delta_{uu'} \delta_{vv'} - \frac{\lambda_c}{K^2}. \quad (9)$$

The negative off-diagonal terms $-\lambda_c/K^2$ explicitly encode competition: increasing one weight reduces the influence of others within the same channel. Crucially, the modulation strength is independently determined by each λ_c , yielding a heterogeneous inhibition landscape that balances sharpening with stability and prevents feature collapse.

Why not Softmax? One might consider applying softmax to enforce competition among kernels. However, it is incompatible with convolution. The simplex probability constraint of softmax removes the essential negative weights for detecting

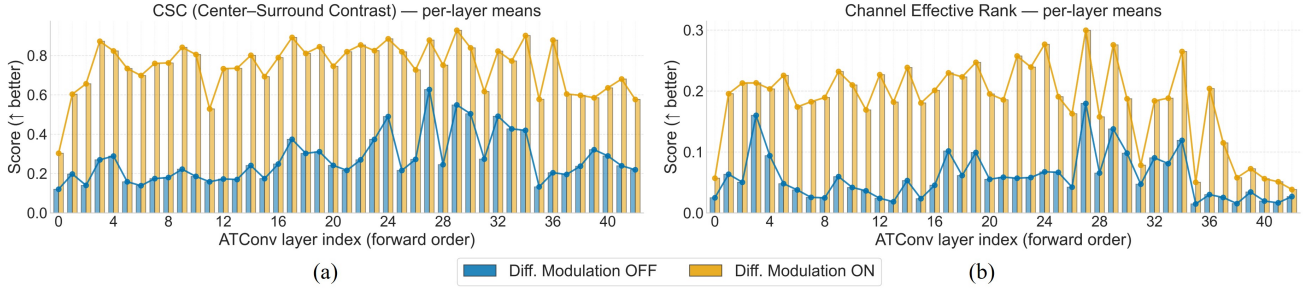


Fig. 3. Impact of Differential Kernel Modulation (Diff. Modulation) on feature properties. We analyze the output of every ATConv in AttNet-T4 with and without using Diff. Modulation. (a) **Center-Surround Contrast** is defined as $CSC(x) = \mathbb{E}[|x - G_\sigma(x)|] \cdot \mathbb{E}[|x|]^{-1}$, where $G_\sigma(\cdot)$ is Gaussian blurring. CSC quantifies the relative enhancement of local pattern contrast, higher means sharper representations. (b) **Channel Effective Rank** is defined as $CER(x) = \exp(H(p)) \cdot C^{-1}$, where $p_i = \lambda_i / \sum_j \lambda_j$ are normalized eigenvalues of the channel covariance and $H(p) = -\sum_i p_i \log p_i$. CER measures the intrinsic dimensionality of channel activations normalized by channel count, indicating diversity of representations, higher means better diversity and less flat response. Both metrics show consistent improvement across all layers with differential kernel modulation, demonstrating enhanced feature sharpness and diversity.

edges and textures, while the exponential scaling drives an extreme winner-take-all effect. It will collapse receptive fields into sparse activations, destroying spatial coherence and leading to training collapse (see Tab. IX-(c)).

In contrast, DKM provides the competitive dynamic tailored for Conv. By preserving signed responses, it enables the simultaneous modeling of excitatory and inhibitory patterns crucial for visual contrast. Its differential form ensures smooth gradient flow without saturation, while its adaptive modulation (λ) automatically discovers heterogeneous suppression strategies across the feature hierarchy: some channels sharpen local discriminative cues, while others capture broader semantic dependencies. Fig. 3 empirically confirms these effects: DKM enhances representation contrast and increases channel effective rank, jointly yielding sharper and less redundant representations. Together, these properties allow convolution to achieve the competitive dynamics while retaining efficiency. **Complete the ATConv Architecture.** With the kernel α in hand, we now complete the ATConv operator. As established by Eq. 4, the adaptive weight alone is not sufficient: if the value space remains fixed, the routing capacity is intrinsically constrained. We therefore leverage a learnable value projection that defines a task-specific basis for ATConv aggregation.

$$\mathbf{V}_{b,c,h,w} = \sum_{i=1}^C \mathbf{W}_{\text{value}}^{(c,i)} \mathbf{X}_{b,i,h,w}. \quad (10)$$

This projection creates a new, task-optimized feature space (determining what to aggregate) where similar concepts cluster, enabling adaptive routing to perform a far more effective and discriminative aggregation.

In summary, ATConv operates in an efficient depth-wise convolutional framework as follows with the proposed C2K, DKM, and value adaptations.

$$\mathbf{Y}_{b,c,h,w}^{\text{ATConv}} = \sum_{u=0}^{K-1} \sum_{v=0}^{K-1} \alpha_{b,c,u,v}^{\text{ATConv}}(\mathbf{X}) \mathbf{V}_{b,c,h+u-p, w+v-p}, \quad (11)$$

where $p = \lfloor K/2 \rfloor$. Both the routing weights α and the feature bases $\mathbf{V}$ adapt to input content under principled rules, allowing ATConv to realize adaptive routing with high efficiency. In contrast to Conv’s static kernels and self-attention’s costly quadratic maps, ATConv demonstrates that global adaptivity

Algorithm 1 Attentive Convolution (ATConv)

Input: Input $\mathbf{X} \in \mathbb{R}^{B \times C \times H \times W}$; kernel size K
Parameters: $\text{Conv}_{1 \times 1}$, $\mathbf{W}_{\text{gen}}$, γ_c , GELU function $\phi(\cdot)$ and Sigmoid function $\sigma(\cdot)$, $\mathbf{W}_{\text{value}}$, $\mathbf{W}_{\text{out}}$.
Output: $\mathbf{Y}^{\text{out}} \in \mathbb{R}^{B \times C \times H \times W}$

// Step-1: Context-to-Kernel Translation
1: $\mathbf{F} \leftarrow \text{Conv}_{1 \times 1}(\mathbf{X})$ # point-wise conv, $\mathbf{F} \in \mathbb{R}^{B \times C \times H \times W}$
2: $\mathbf{Z} \leftarrow \text{AdaAvgPool}_{K \times K}(\mathbf{F})$ # $\mathbf{Z} \in \mathbb{R}^{B \times C \times K \times K}$
3: $\mathbf{K} \leftarrow \text{Reshape}(\mathbf{W}_{\text{gen}} \cdot \text{Vec}(\phi(\mathbf{Z})))$ # $\mathbf{K} \in \mathbb{R}^{B \times C \times K \times K}$
// Step-2: Lateral inhibition via Diff. Kernel Modulation
4: $\bar{\mathbf{K}}_{b,c} \leftarrow \text{mean}(\mathbf{K}_{b,c,:,:})$ # spatial mean per channel
5: $\alpha_{b,c,:,:} \leftarrow \mathbf{K}_{b,c,:,:} - \sigma(\gamma_c) \bar{\mathbf{K}}_{b,c}$ # per channel competition
// Step-3: Convolution with adaptive routing
6: $\mathbf{V} \leftarrow \mathbf{W}_{\text{value}} \cdot \mathbf{X}$ # value projection
7: $\mathbf{Y} \leftarrow \alpha \star \mathbf{V}$ # depthwise conv process
// Step-4: Output projection
8: $\mathbf{Y}^{\text{out}} \leftarrow \mathbf{W}_{\text{out}} \cdot \mathbf{Y}$ # output projection
9: **return** $\mathbf{Y}^{\text{out}}$

can be achieved within a compact local processor, redefining how convolutional architectures can embody the expressive power once thought to be exclusive to attention. In the following narrative, we use the term “ $\star$ ” to denote the depthwise convolutional operation in Eq. 11 for simplicity.

Based on the above designs, we define ATConv in Alg. 1. For visual clarity, its architecture is illustrated in Fig. 2-(b) and (c). Note that following self-attention, we also use an additional linear projection ($\mathbf{W}_{\text{out}}$) before final output.

D. Complexity Analysis

ATConv preserves the dynamic expressivity of self-attention while being substantially more efficient. We quantify these gains from two complementary viewpoints: computational complexity and memory footprint.

Computational Complexity. For $\mathbf{X} \in \mathbb{R}^{B \times C \times H \times W}$ with $N = HW$, vanilla self-attention has quadratic complexity:

$$\mathcal{O}_{\text{SA}} = \underbrace{\mathcal{O}(NC^2)}_{\text{projections}} + \underbrace{\mathcal{O}(N^2C)}_{\text{attention map}}, \quad (12)$$

Instead, ATConv consists of (i) context-to-kernel translation,

(ii) convolutional aggregation, and (iii) linear projections:

$$\begin{aligned} \mathcal{O}_{\text{ATConv}} &= \underbrace{\mathcal{O}(NC^2)}_{\text{context} \rightarrow \text{kernel}} + \underbrace{\mathcal{O}(NK^2C)}_{\text{conv}} + \underbrace{\mathcal{O}(NC^2)}_{\text{projections}} \\ &= \mathcal{O}(NC(C+K^2)) \approx \mathcal{O}(NC^2), \end{aligned} \quad (13)$$

typically $K^2 \ll C$. Thus, ATConv is *linear in N* , avoiding the quadratic bottleneck of self-attention.

Memory Footprint Analysis. Beyond arithmetic complexity, memory is the practical bottleneck for modern deep learning models. Self-attention stores an $N \times N$ attention map (with softmax buffers), yielding a quadratic term, whereas ATConv replaces it with a compact bank of dynamic kernels:

$$\begin{aligned} \text{Mem}_{\text{SA}} &= \mathcal{O}(BNC) + \mathcal{O}(BN^2), \\ \text{Mem}_{\text{ATConv}} &= \mathcal{O}(BNC) + \mathcal{O}(BCK^2). \end{aligned} \quad (14)$$

Furthermore, the routing branch of ATConv is consumed on the fly, so no full BNC activation is persisted; only one large activation (e.g., V or X) plus a compact BCK^2 kernel buffer is retained. By contrast, in addition to the huge $N \times N$ attention maps, SA needs extra $3BNC$ buffers for the three activations Q , K , and V that must be cached for backpropagation. We give a practical example to illustrate the actual memory footprint differences. Generally, let $B=32$, $C=384$, $H=W=28$ ($N=784$), $K=3$, FP16 (2 bytes/elt). Using the dominant terms in (14) with cached $Q/K/V$ will have the following cost:

$$\begin{aligned} \text{Mem}_{\text{SA}} &\approx 2 \left(\underbrace{3BNC}_{Q, K, V} + \underbrace{BN^2}_{\text{attn map}} \right) \approx 92.6 \text{ MiB}, \\ \text{Mem}_{\text{ATConv}} &\approx 2 \left(\underbrace{BNC}_{\text{acts}} + \underbrace{BCK^2}_{\text{dyn kernels}} \right) \approx 18.6 \text{ MiB}. \end{aligned} \quad (15)$$

Under this accounting, ATConv reduces per-layer memory by 79.9% at 28×28 and by 95% at 56×56 , reflecting the removal of the quadratic BN^2 term. In short, SA's $\mathcal{O}(BN^2)$ buffer is replaced by ATConv's channel-local $\mathcal{O}(BCK^2)$ buffer, yielding considerable savings while preserving attention-like global dynamics. Even with FlashAttention, which streams computation and avoids storing the explicit $N \times N$ map, narrowing peak memory to $\mathcal{O}(BNC)$, its token2token interactions and memory traffic remain quadratic across tiles. In contrast, ATConv stays linear in N and preserves cache-friendly locality with a $\mathcal{O}(BCK^2)$ working set, yielding higher throughput and lower memory consumption.

E. Implementation Details

Overall Architecture of AttNet. Based on ATConv, we construct AttNet as a family of general-purpose visual backbones, as shown in Fig. 2. Specifically, we adopt the ATConv as the spatial operator (token-mixer) and follow the modern ViT architecture to build the model. Following recent best practices [17], [50]–[52], we employ the Gated Linear Units (GLU) as a lightweight alternative to the traditional Feed-Forward Network (FFN) for channel mixing. All other components (e.g., double skip connections, Layer Normalization, GELU activation) remain consistent with ViT styles. In this manner, we develop four variants of AttNet with different budgets,

TABLE I
CONFIGURATION OF FOUR ATTNET VARIANTS. THE NUMBER OF BLOCKS AND CHANNELS ARE CONFIGURED FOR FOUR STAGES.

Model	#Params	FLOPs	#Blocks	#Channels
AttNet-T1	13.7M	2.4G	(2, 3, 12, 3)	(48, 96, 224, 384)
AttNet-T2	27.0M	5.1G	(3, 3, 16, 3)	(64, 128, 288, 512)
AttNet-T3	49.1M	9.4G	(4, 4, 26, 4)	(72, 144, 320, 576)
AttNet-T4	87.3M	16.7G	(5, 5, 28, 5)	(96, 192, 384, 768)

denoted AttNet-T1, -T2, -T3, and -T4. The model sizes and configurations are listed in Tab. I.

ATConv Configuration. As discussed in Sec. III-B, our key insight is to inject adaptive routing and lateral inhibition to achieve higher expressivity, rather than simply enlarging the receptive field. Although using larger kernels may yield further improvements, we follow our findings and *configure ATConv with a compact 3×3 kernel size for all AttNet variants.*

IV. EXPERIMENTS

In this section, we first show that ATConv can serve as a drop-in replacement for strong SA mechanisms in image classification, delivering a superior accuracy–efficiency trade-off (Sec. IV-A). We then compare ATConv with state-of-the-art SA variants in computational efficiency (Sec. IV-B). Next, we evaluate AttNet across core vision tasks, including image classification (Sec. IV-C), object detection (Sec. IV-D), and semantic segmentation (Sec. IV-E). We additionally assess robustness of ATConv on cross-modal retrieval (Sec. IV-F) and examine the utility of ATConv in diffusion-based image generation (Sec. IV-G). Finally, we present ablations and analyze architectural design choices (Sec. IV-H).

A. ATConv as a Drop-in Replacement for Self-Attention

To rigorously assess whether ATConv can replace self-attention in modern vision backbones, we perform controlled “drop-in replace” experiments on two canonical ViTs: PVT [12] and Swin [9] Transformer, representing pooling-based and window-based visual self-attention designs, respectively. All experiments are conducted on ImageNet-1K under identical training protocols and architectural settings for both baselines and their ATConv replaced variants. We substitute their self-attention modules with ATConv blocks. Besides, we replace the classification token with global average pooling for final feature aggregation, since ATConv's dense spatial processing renders a dedicated classification token unnecessarily.

Tab. II reports results across multiple model scales. For the PVT family, ATConv delivers substantial gains: PVT-Tiny with ATConv achieves 2.4% higher Top-1 accuracy while improving the throughput by $1.5\times$. The Swin Transformer family shows equally compelling results. On the base scale, ATConv-Swin-B improves Top-1 accuracy by 0.8% while achieving a speedup over $2\times$, confirming the effectiveness of ATConv in large scales where self-attention typically excels over traditional operators. Across all Swin variants (Tiny through Base), ATConv consistently surpasses window attention in both accuracy and efficiency, with speedups ranging from $1.8\times$ to $2.2\times$. By increasing the resolution to 384px ,

TABLE II

COMPARISON WITH BASELINES ON IMAGENET-1K DATASET. WE REPLACE THE ATTENTION MECHANISM IN PVT [12] AND SWIN [9] WITH ATCONV IN A DROP-IN MANNER TO SHOW THE VARIATION IN ACCURACY AND SPEED. THE THROUGHPUT (THP.) METRIC IS MEASURED ON ONE MI-250X GPU.

Method	#Params (M)	FLOPs (G)	Thp. (fps) ↑	Top-1 (%) ↑
PVT-T [12]	13.2	1.9	1701	75.1
ATConv-PvT-T	10.2	1.8	2476 (1.5×)	77.5 (+2.4)
PVT-S [12]	24.5	3.8	939	79.8
ATConv-PvT-S	18.3	3.5	1479 (1.6×)	81.7 (+1.9)
PVT-M [12]	44.2	6.7	590	81.2
ATConv-PvT-M	31.9	6.1	879 (1.5×)	82.4 (+1.2)
PVT-L [12]	61.8	9.8	421	81.7
ATConv-PvT-L	43.3	9.2	662 (1.6×)	83.0 (+1.3)

Method	#Params (M)	FLOPs (G)	Thp. (fps) ↑	Top-1 (%) ↑
Swin-T [9]	28.3	4.5	958	81.3
ATConv-Swin-T	28.0	4.2	1748 (1.8×)	82.2 (+0.9)
Swin-S [9]	49.6	8.8	539	83.0
ATConv-Swin-S	47.6	8.1	967 (1.8×)	83.6 (+0.6)
Swin-B [9]	87.8	15.5	364	83.5
ATConv-Swin-B	84.2	14.3	789 (2.2×)	84.3 (+0.8)
Swin-B-384 [9]	87.9	47.2	105	84.5
ATConv-Swin-B-384	84.3	42.1	329 (3.1×)	85.0 (+0.5)

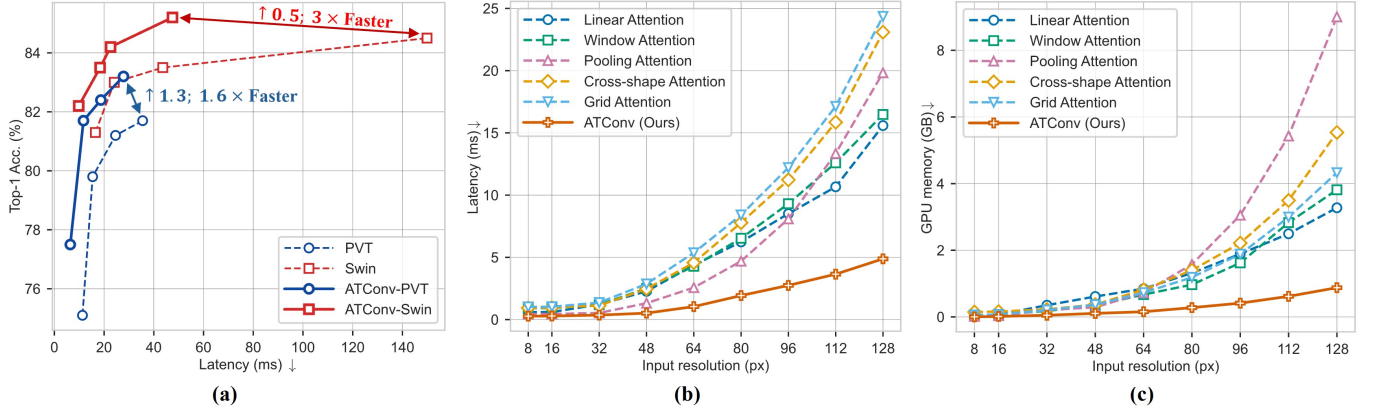


Fig. 4. (a): Latency vs. accuracy curve on PVT [12] and Swin transformers [9], with their ATConv versions which directly replace the self-attention blocks with ATConv; (b) and (c): Latency and GPU memory consumption in terms of different input resolutions on different operators at the atomic level.

ATConv can bring further accelerations by $3.1\times$ with a consistent accuracy gain of 0.5%. Fig. 4-(a) further illustrates the latency–accuracy trade-off, where ATConv consistently reduces inference latency while delivering accuracy gains.

The consistent improvements across diverse model scales indicate that ATConv captures fundamental visual structures more efficiently than representative self-attention variants in modern ViTs. The results establish ATConv as a principled convolutional alternative for visual self-attention that advances both accuracy and computational efficiency.

B. Operator-Level Efficiency Comparison

To assess operator-level efficiency across varying input sizes, we benchmark ATConv against five leading attention mechanisms representing the state-of-the-art accuracy–efficiency trade-off: (1) InLine Linear Attention [31], (2) Cross-Shaped Attention from Cswin [11], (3) Improved pooling Attention from PVTv2 [16], (4) Window Attention from Swin [9], and (5) Grid Attention from MaxViT [30]. For a controlled comparison, each operator is implemented as a standalone module and evaluated directly on input tensors without down-sampling. We fix the channel dimension at $C = 128$ and batch size at $B = 64$, while progressively increasing spatial resolution from 8×8 to 128×128 . Latency (ms) and peak GPU memory (MB) are measured on a single MI-250X GPU, with each data point averaged over 10 runs.

Latency. As shown in Fig. 4-(b), ATConv consistently outperforms all baselines across resolutions. At the input resolution of 128px, ATConv is $2.1\times$ faster than the InLine Linear

Attention and $4.5\times$ faster than Grid Attention. These results demonstrate that ATConv scales more gracefully with spatial resolution, avoiding the quadratic or fragmented computation patterns that burden attention mechanisms. This operator-level evidence confirms that ATConv offers fundamentally higher computational efficiency, particularly on large-resolution inputs where efficiency bottlenecks are most critical.

Memory footprint. Fig. 4-(c) shows greater advantages about memory consumption: ATConv consumes only 1/3 to 1/10 of the peak memory required by the compared attention mechanisms at 128px. This huge reduction stems from its elimination of key–query intermediate tensors and the absence of large attention maps, which dominate memory usage in attention. These statistics underscore ATConv’s hardware friendliness, allowing deployment in memory-constrained scenarios such as on edge devices, with improved training stability on large-scale servers.

Summary. By jointly reducing latency and memory pressure, ATConv provides a compelling efficiency–accuracy balance. Its favorable scaling properties and hardware adaptability establish it as a principled and practical alternative to visual attention, particularly for applications demanding both high throughput and low resource consumption.

C. Image Classification

Image classification is a fundamental computer vision task, where the goal is to assign a class label to each input image. Many other tasks (e.g., detection and segmentation) build upon networks pretrained on classification as feature extractors.

TABLE III

IMAGE CLASSIFICATION RESULTS ON THE IMAGENET-1K DATASET. TOP-1 INDICATES THE TOP-1 ACCURACY. FLOPS METRICS ARE MEASURED WITH THE INPUT RESOLUTION OF 224×224 .

Method	Type	#Params (M)	FLOPs (G)	Top-1 (%)
PVTv2-B1 [16]	ViT	13.1	2.1	78.7
BiFormer-T [53]	ViT	13.1	2.2	81.4
EfficientFormerV2-S2 [54]	ViT	12.7	1.3	82.0
TransNeXt-Micro [17]	ViT	12.8	2.7	82.5
AttNet-T1	CNN	13.7	2.4	82.8
Swin-T [9]	ViT	28.3	4.5	81.3
PVTv2-B2 [16]	ViT	25.4	4.0	82.1
Focal-T [55]	ViT	29.1	4.9	82.2
EfficientFormerV2-L [54]	ViT	26.1	5.2	83.5
STViT-S [56]	ViT	25.0	4.4	83.6
MaxViT-Tiny [30]	ViT	31.0	5.6	83.6
BiFormer-S [53]	ViT	25.5	4.5	83.8
InLine-CSwin-S [31]	ViT	33.0	4.3	83.8
TransNeXt-Tiny [17]	ViT	28.2	5.7	84.0
AttNet-T2	CNN	27.0	5.1	84.4
Swin-S [9]	ViT	49.6	8.7	83.0
PVTv2-B3 [16]	ViT	45.2	6.9	83.2
Focal-S [55]	ViT	51.1	9.1	83.5
PVTv2-B4 [16]	ViT	62.6	10.1	83.6
BiFormer-B [53]	ViT	56.8	9.8	84.3
MaxViT-S [30]	ViT	68.9	11.7	84.5
TransNeXt-Small [17]	ViT	49.7	10.3	84.7
STViT-B [56]	ViT	52.0	9.9	84.8
AttNet-T3	CNN	49.1	9.4	85.3
Swin-B [9]	ViT	87.8	15.4	83.5
PVTv2-B5 [16]	ViT	82.0	11.8	83.8
Focal-B [55]	ViT	89.8	16.0	83.8
InLine-CSwin-B [31]	ViT	73.0	14.9	84.5
TransNeXt-Base [17]	ViT	89.7	18.4	84.8
MaxViT-B [30]	ViT	120.0	23.4	84.9
STViT-L [56]	ViT	95.0	15.6	85.3
AttNet-T4	CNN	87.3	16.7	85.6

Here, we evaluate AttNet on the ImageNet-1K [60] dataset and compare with state-of-the-art ViTs.

Experimental Setup. For a fair comparison, we follow the widely accepted protocols in DeiT [8] to train and evaluate our model on the ImageNet-1K dataset. Briefly, AttNet is trained from scratch on ImageNet-1K for 300 epochs, with a total batch size of 4096 distributed across 64 AMD MI-250X GPUs. We use the AdamW [61] optimizer with a peak learning rate of $4e-3$ and a weight decay of 0.05. A 5-epoch linear warm-up is followed by a cosine decay schedule to $1e-5$. All training and testing images are resized to 224×224 . We adopt commonly accepted augmentations used in DeiT and many other ViTs [14], [16]–[18], including RandAugment, MixUp, CutMix, and random erasing. All settings remain consistent for classification across ablations unless explicitly stated.

Experimental Results. The quantitative results are summarized in Tab. III. For speed comparison, we provide operator-level measurements in Figs. 4-(b) and -(c), focusing on the top-5 fastest attention mechanisms selected in Tab. III. Thus, we omit network-level speed metrics here, as they are often confounded by additional architectural factors (e.g., activation functions, depth, and width). In contrast, the atomic operator-level benchmark provided in Fig. 4 offers a more faithful

TABLE IV

RESULTS ON COCO VAL2017 DATASETS BY DROP-IN REPLACING THE PVT’S POOLING ATTENTION WITH ATCONV. FLOPS AND THROUGHPUT (THP.) ARE MEASURED UNDER 1280×800 RESOLUTION.

Mask R-CNN Object Detection on COCO (1×)							
Backbone	FLOPs (G)	Thp. (fps)	AP^b	AP_{50}^b	AP_{75}^b	AP^m	AP_{50}^m
PVT-T [16]	240	54	36.7	59.2	39.3	35.1	56.7
InLine-PVT-T [31]	211	79	40.2	62.7	43.8	37.7	59.7
ATConv-PVT-T	173	104	42.3	64.2	45.6	39.3	62.2
PVT-S [16]	305	29	40.4	62.9	43.8	37.7	59.7
InLine-PVT-S [31]	250	47	43.4	66.4	47.1	40.1	63.1
ATConv-PVT-S	215	70	45.9	68.5	49.6	43.3	65.7
PVT-M [16]	392	18	42.0	64.4	45.6	39.0	61.6
InLine-PVT-M [31]	301	28	44.0	66.4	48.0	40.3	63.4
ATConv-PVT-M	252	45	46.2	68.8	50.3	42.7	65.8
PVT-L [16]	494	12	42.9	65.0	46.6	39.5	61.9
InLine-PVT-L [31]	377	17	45.4	67.6	49.7	41.4	64.7
ATConv-PVT-L	332	32	47.3	69.8	51.8	43.3	66.5

assessment of the intrinsic efficiency.

As shown in Tab. III, AttNet clearly outperforms state-of-the-art ViTs such as TransNeXt [17] and STViT [56]. For instance, AttNet-T1/T2/T3 achieve Top-1 accuracy improvements of 1.4%, 0.6%, and 1.0% over BiFormer, respectively. Compared with recent ViTs that combine sophisticated attention mechanisms with DWConvs, AttNet demonstrates notable advantages while entirely removing the reliance on attention. In particular, AttNet-T2/T3/T4 exceed TransNeXt-Tiny/Small/Base by 0.4%, 0.6%, and 0.8% Top-1 accuracy, respectively, with fewer parameters and FLOPs.

Building upon ATConv, AttNet operates with a purely convolutional architecture to surpass state-of-the-art ViTs. These results demonstrate that by embedding the core advantages of attention into convolutional design, convolutional operators can not only match but also exceed the performance of attention-based models while offering superior efficiency.

D. Object Detection and Instance Segmentation

Object detection and instance segmentation have long been fundamental and challenging tasks in computer vision. These tasks aim to detect and recognize instances of semantic objects within natural images. In this section, we evaluate AttNet on the MS-COCO [62] dataset.

Experimental Setup. Following common practices [9], [31], [55], [58] in the community, we utilize pretrained models on ImageNet-1K as the backbone, integrating Mask R-CNN [63] and Cascaded R-CNN [64] as the detection and segmentation heads. The models are fine-tuned on the MS-COCO dataset using the AdamW optimizer, following two common experimental configurations: “1×” (12 training epochs) and “3×+MS” (36 training epochs with multi-scale training). For comparative analysis, we use the configurations from Swin Transformer and Cswin. All training and evaluations are conducted with MMDetection on a distributed setup using 64 AMD MI-250X GPUs.

Drop-in Evaluation. Tab. IV summarizes the experimental results for the drop-in replacement of pooling attention mechanisms in PVT with our proposed ATConv. Our ATConv-PVT

TABLE V

OBJECT DETECTION AND INSTANCE SEGMENTATION WITH MASK R-CNN ON COCO VAL2017 DATASET. THE FLOPS ARE MEASURED AT RESOLUTION 800×1280 . ALL MODELS ARE PRETRAINED ON IMAGENET-1K. MS SCHEDULE MEANS RESULTS WITH MULTI-SCALE TRAINING.

Backbone	FLOPs (G)	Type	Mask R-CNN $1 \times$						Mask R-CNN $3 \times$ + MS schedule					
			AP^b	AP_{50}^b	AP_{75}^b	AP^m	AP_{50}^m	AP_{75}^m	AP^b	AP_{50}^b	AP_{75}^b	AP^m	AP_{50}^m	AP_{75}^m
Swin-T [9]	264	ViT	42.2	64.4	46.2	39.1	61.6	42.0	46.0	68.2	50.2	41.6	65.1	44.8
Focal-T [55]	291	ViT	44.8	67.7	49.2	41.0	64.7	44.2	47.2	69.4	51.9	42.7	66.5	45.9
CMT-S [57]	249	ViT	44.6	66.8	48.9	40.7	63.9	43.4	48.3	70.4	52.3	43.7	67.7	47.1
UniFormer-S [58]	269	ViT	45.6	68.1	49.7	41.6	64.8	45	48.2	70.4	52.5	43.4	67.1	47.0
AttNet-T2	225	CNN	47.3	69.2	51.8	42.6	66.3	45.9	48.6	70.6	54.0	44.2	67.5	47.4
Swin-S [9]	354	ViT	44.8	66.6	48.9	40.9	63.4	44.2	48.5	70.2	53.5	43.3	67.3	46.6
Focal-S [55]	401	ViT	47.4	69.8	51.9	42.8	66.6	46.1	48.8	70.5	53.6	43.8	67.7	47.2
DAT-S [59]	378	ViT	47.1	69.9	51.5	42.5	66.7	45.4	49	70.9	53.8	44.0	68.0	47.5
UniFormer-B [58]	399	ViT	47.4	69.7	52.1	43.1	66.0	46.5	50.3	72.7	55.3	44.8	69.0	48.3
AttNet-T3	329	CNN	49.6	71.3	54.2	44.5	68.6	48.3	50.7	72.2	55.5	45.4	69.4	49.2
Swin-B [9]	496	ViT	46.9	—	—	42.3	—	—	48.5	69.8	53.2	43.4	66.8	46.9
Cswin-B [11]	526	ViT	48.7	70.4	53.9	43.9	67.8	47.3	50.8	72.1	55.8	44.9	69.1	48.3
AttNet-T4	478	CNN	51.3	72.6	56.5	45.6	69.8	49.6	51.6	72.9	56.8	45.7	70.2	49.9

TABLE VI

RESULTS OF SEMANTIC SEGMENTATION BY DROP-IN REPLACING THE ATTENTION IN PVT SERIES WITH ATCONV. FLOPS ARE MEASURED WITH AN INPUT SPATIAL SIZE OF 512×2048 . THROUGHPUT (THP.) METRICS ARE MEASURED ON ONE MI-250X GPU.

Semantic Segmentation on ADE20K					
Backbone	FLOPs (G)	Params (M)	Thp. (fps)	mIoU (%) $\uparrow$	mAcc (%) $\uparrow$
PVT-T [12]	158	17	54	36.57	46.72
InLine-PVT-T [31]	127	16	74	39.16	50.63
ATConv-PVT-T	136	14	98	42.43	53.89
PVT-S [12]	225	28	20	41.95	53.02
InLine-PVT-S [31]	168	25	39	42.93	54.58
ATConv-PVT-S	173	23	60	45.81	56.60
PVT-L [12]	420	65	12	43.99	54.62
InLine-PVT-L [31]	298	55	16	44.71	57.17
ATConv-PVT-L	292	48	29	48.32	59.18

models significantly outperform the baseline PVT models, demonstrating substantial improvements in both object detection and instance segmentation tasks. Specifically, replacing the pooling attention in PVT-T with ATConv leads to a notable increase in average precision (AP^b) from 36.7 to 42.3, while throughput (Thp.) improves from 54 fps to 104 fps. Similarly, for larger PVT variants like PVT-L, ATConv-PVT-L achieves an AP^b improvement from 42.9 to 47.3, alongside a throughput boost from 12 fps to 32 fps. These results underscore the efficacy of ATConv in enhancing both performance and efficiency compared to the standard pooling-based attention mechanisms in vision transformers.

We further compare our ATConv-based approach with InLine attention [31], a state-of-the-art linear attention mechanism. ATConv-PVT consistently outperforms its InLine counterparts with between accuracy and efficiency. For instance, ATConv-PVT-T achieves an AP^b of 42.3, surpassing InLine-PVT-T's AP^b of 40.2, with a significant throughput improvement from 79 fps (InLine-PVT-T) to 104 fps. This trend holds across other variants: ATConv-PVT-S outperforms InLine-PVT-S with an AP^b of 45.9 versus 43.4, and ATConv-PVT-M

similarly exceeds InLine-PVT-M (46.2 vs. 44.0 at AP^b).

Comparison with State-of-the-Art Models. Tab. V compares ATConv-based backbones with several state-of-the-art models, including Focal Transformer [55], UniFormer [58], DAT [59], and Cswin [11]. All the compared methods leverage strong attention mechanisms that are powerful for dense prediction. Our AttNet consistently outperforms these models across all metrics with efficient 3×3 spatial kernels. Specifically, AttNet-T2 achieves an AP^b of 47.3, surpassing Focal-T (44.8) and UniFormer-S (45.6). Additionally, AttNet-T4 outperforms Cswin-B, with a 2.6-point higher AP^b and a 1.7-point higher AP^m under the $1 \times$ training schedule.

These results highlight the competitive performance of AttNet against ViTs in dense prediction tasks. In particular, with only a 3×3 kernel, AttNet achieves performance on par with various attention variants with larger receptive fields. This underscores the key contribution of our ATConv, which efficiently encodes global scene understanding into local processing rules with the context-to-kernel translation. It provides an effective and computationally efficient alternative to attention mechanisms for object understanding.

E. Semantic Segmentation

Semantic segmentation involves assigning a semantic label to each pixel in an image, making it one of the most critical tasks in computer vision that asserts the dense prediction capacity of foundation models. We benchmark the proposed ATConv on the ADE20K dataset for semantic segmentation.

Setup. We evaluate the performance of ATConv when integrated into the PVT series for semantic segmentation on the ADE20K [70] dataset. We replace the attention mechanism in the PVT backbone with the proposed ATConv and measure the performance in terms of mean Intersection over Union (mIoU) and mean Accuracy (mAcc). We employ SemanticFPN as segmentation heads and follow the protocols in [12], [31] for fair comparisons.

Experimental results. As shown in Tab. VI, the results indicate that ATConv consistently outperforms both the PVT and InLine variants in terms of segmentation accuracy, while

TABLE VII
COMPARISON ON LLCM [65] AND VCM-HITSZ [66] BENCHMARKS FOR VISIBLE-INFRARED IMAGE RETRIEVE IN THE CROSS-MODALITY SETTING. r DENOTES RANK- k ACCURACY, mAP DENOTES MEAN AVERAGE PRECISION, AND mINP DENOTES MEAN INVERSE NEGATIVE PENALTY.

Method	Params (M)	LLCM [65]					VCM-HITSZ [66]				
		$r=1$	$r=5$	$r=20$	mAP	mINP	$r=1$	$r=5$	$r=20$	mAP	mINP
EfficientFormerV2-S2 [54]	12.6	42.51	66.49	84.88	50.15	46.62	34.71	54.14	71.31	24.14	7.88
PVTv2-B1 [16]	13.1	48.36	71.43	88.02	55.71	52.37	51.35	68.61	81.26	37.45	15.24
AttNet-T1	13.7	50.85	73.28	88.67	57.82	54.49	56.50	72.77	83.51	42.47	18.00
ResNet-50 [1]	25.6	36.47	59.36	79.21	43.54	39.80	37.30	56.20	71.75	23.54	5.71
ConvNeXt-Tiny [21]	29.0	36.56	60.15	79.76	44.04	40.62	5.04	14.38	29.56	3.76	0.83
Swin-Tiny [9]	29.0	42.06	66.90	85.32	50.10	46.80	51.55	67.56	79.97	38.55	16.70
EfficientFormerV2-L [54]	26.1	45.83	69.39	86.57	53.25	49.75	46.62	63.53	77.53	33.98	13.30
Focal-Tiny [55]	29.1	48.55	71.38	87.73	55.72	52.35	59.60	75.20	85.17	46.42	21.21
PVTv2-B2 [16]	25.4	48.82	71.56	87.80	56.03	52.70	58.60	73.73	84.25	44.07	19.99
MaxViT-Tiny [30]	31.0	50.00	72.62	88.41	57.15	53.79	59.60	74.15	84.18	47.09	22.50
STViT-Small [56]	25.0	50.18	72.78	88.65	57.06	53.51	54.08	70.09	82.22	41.56	18.35
AttNet-T2	27.0	52.41	74.85	89.60	59.54	56.32	62.74	76.07	85.75	50.07	25.46

TABLE VIII
FID COMPARISONS WITH VANILLA SiT AND REPA ON FFHQ [67] AT 512×512 RESOLUTION AND IMAGENET-1K [67] AT 256×256 RESOLUTION. FOR FFHQ, WE DO NOT USE CLASSIFIER-FREE GUIDANCE (CFG) AND SAMPLING USING UNCONDITIONAL GENERATION SETTINGS. FOR IMAGENET-1K, WE SAMPLING USE A CONSISTENT CFG SCALE OF 1.8 WITHOUT ADDITIONAL SCHEDULING. $\downarrow$ DENOTES LOWER THE BETTER.

Model	#Params (M)	Iter.	Lat. $\downarrow$	FFHQ [67] 512×512				ImageNet-1K [60] 256×256 with REPA [68]					
				FID $\downarrow$	sFID $\downarrow$	Pre. $\uparrow$	Rec. $\uparrow$	Lat. $\downarrow$	FID $\downarrow$	sFID $\downarrow$	IS $\uparrow$	Pre. $\uparrow$	Rec. $\uparrow$
SiT-B/2 [69]	130.32	400K	139.64	10.42	26.45	0.60	0.47	31.17	8.01	5.78	147.72	0.70	0.57
SiT-Hybrid-B/2	126.78	400K	112.43	10.09	19.76	0.62	0.53	27.41	7.17	5.44	153.78	0.72	0.59
SiT-ATConv-B/2	123.24	400K	103.14	10.31	23.76	0.63	0.52	25.45	7.15	5.10	149.23	0.73	0.58
SiT-L/2 [69]	457.84	400K	412.71	9.36	24.75	0.62	0.51	97.44	2.12	4.87	265.50	0.79	0.56
SiT-Hybrid-L/2	445.21	400K	375.19	7.74	16.83	0.65	0.57	89.71	1.97	4.66	262.28	0.81	0.60
SiT-ATConv-L/2	432.68	400K	341.01	7.99	16.13	0.66	0.59	82.22	1.95	4.61	267.75	0.81	0.61
SiT-XL/2 [69]	674.83	400K	647.75	8.87	18.48	0.65	0.56	139.44	1.97	4.76	282.33	0.79	0.58
SiT-Hybrid-XL/2	656.26	400K	565.96	7.72	17.91	0.67	0.60	121.52	1.86	4.76	290.61	0.82	0.61
SiT-ATConv-XL/2	637.68	400K	501.39	7.88	17.12	0.69	0.61	113.05	1.82	4.71	291.17	0.83	0.62

also achieving superior throughput. For example, ATConv-PVT-T improves the mIoU to 42.43% and mAcc to 53.89%, surpassing the InLine-PVT-T model, which achieves a mIoU of 39.16% and mAcc of 50.63%. Notably, ATConv-PVT-T achieves a remarkable throughput of 98 fps, significantly higher than both PVT-T (54 fps) and InLine-PVT-T (74 fps).

In the larger model configurations, ATConv continues to show strong improvements. For instance, ATConv-PVT-S outperforms InLine-PVT-S with a mIoU of 45.81% and mAcc of 56.60%, along with a throughput of 60 fps, compared to 39 fps for InLine-PVT-S. Similarly, ATConv-PVT-L achieves a mIoU of 48.32% and mAcc of 59.18%, with an impressive throughput of 29 fps, outperforming InLine-PVT-L, which has a mIoU of 44.71% and mAcc of 57.17% at 16 fps.

These results highlight the effectiveness of ATConv in enhancing both the segmentation accuracy and efficiency of PVT series, demonstrating its potential as a drop-in replacement of attention mechanisms in semantic segmentation tasks.

F. Evaluation on Cross-domain Robustness

We evaluate ATConv’s robustness on cross-modality understanding using the challenging LLCM [65] and VCM-HITSZ [66] datasets for visible-infrared retrieval. This task requires establishing reliable latent correspondence between visible and infrared modalities under huge imaging spectral discrepancies.

While self-attention typically demonstrates superior representational robustness over convolution, we show that ATConv achieves better robustness to various self-attention variants.

Setup. All methods are trained on LLCM [65] and VCM-HITSZ [66] using ImageNet-1K pretrained models, following the standard image retrieval training and evaluation protocol in [71]. For fair comparison, all images are resized to 224×224 to meet the strict input resolution requirements of some attention-based baselines. Following common practices [72]–[75], we comprehensively report the rank at r , mean Average Precision (mAP), and mean Inverse Negative Penalty (mINP) [71] as evaluation metrics (all metrics are higher the better).

Results. Tab. VII demonstrates AttNet’s superior robustness on cross-modality understanding. AttNet-T1 achieves the mAP of 57.82%/42.47% on LLCM/VCM-HITSZ, surpassing PVTv2-B1 by 2.11%/5.02% with comparable parameters. AttNet-T2 establishes SOTA mAP with 59.54%/50.07% mAP on LLCM/VCM-HITSZ, significantly outperforming the attention-based solutions MaxViT-Tiny (57.15%/47.09%) and STViT-Small (57.06%/41.56%). AttNet also consistently achieves the best rank- r and mINP metrics for feature-based retrieval, demonstrating superior capacity for learning robust representations in heterogeneous latent spaces.

The robustness of AttNet comes from adaptive routing and lateral inhibition, the former adaptively aggregates se-



Fig. 5. Representative images generated by SiT-ATConv-XL/2, where all attention mechanisms are replaced with ATConv to form a pure convolutional generative architecture. The top two rows present natural images synthesized on the ImageNet-1K dataset using classifier-free guidance ($w = 4.0$). The bottom row shows facial images unconditionally synthesized on the FFHQ dataset. We show results at 400K steps, more training steps will yield better quality.

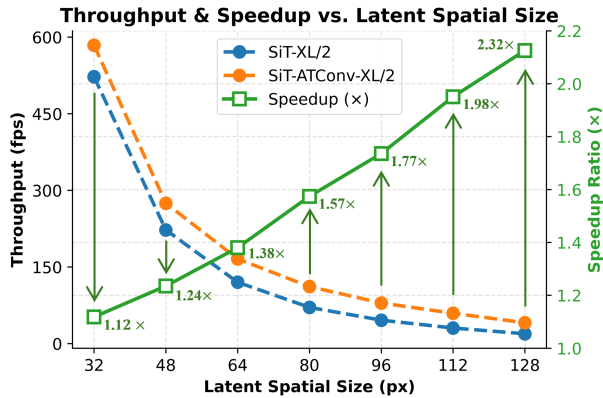


Fig. 6. Comparison of throughput scaling with different latent sizes. ATConv achieves greater efficiency advantage at larger latent resolutions.

manically relevant features across modalities despite spectral gaps, while the latter injects competitive dynamics to suppress noise and amplify modality-shared cues. These mechanisms enable AttNet to achieve better efficiency-robustness tradeoff, particularly evident in the 8.51% improvement in mAP over the robust STViT-Small on the VCM-HITSZ benchmark.

G. ATConv for Diffusion Image Generation

Diffusion models [68], [69], [76], [77] have become the cornerstone of visual AIGC, achieving an unprecedented quality of visual synthesis. Unlike traditional discriminative tasks (e.g., classification and detection) that identify patterns from existing images, diffusion generation demands constructing coherent visual structures from noise, demanding exceptional capacity for both global semantic consistency and fine-grained detail synthesis. This complexity has established transformer [68], [69], [78], [79] as the dominant architecture.

We challenge this assumption by investigating whether ATConv, as an efficient convolutional operator, can match the generation quality of self-attention. Departing from conventional research limited to discriminative tasks, in this paper, we

directly test ATConv in the diffusion-based image generation task where attention has been deemed irreplaceable.

Setup. We evaluate ATConv in two scenarios: (1) unconditional 512×512 facial image generation on FFHQ [67] using SiT [69] as the baseline, and (2) conditional 256×256 natural image generation on ImageNet-1K [60] using SiT [69] with REPA [68] as the baseline. Building upon the baseline, we create SiT-ATConv variants by replacing all attention blocks in SiT with our ATConv. Besides, we further present hybrid variants (SiT-Hybrid) that replace only even-numbered attention blocks with ATConv for an interleaved architecture. All models maintain identical training protocols and evaluation metrics as their baselines [68], [69] to ensure fair comparisons.

Results. Tab. VIII presents quantitative results, revealing that ATConv exceeds self-attention for diffusion-based image generation with better generation quality and efficiency. On FFHQ at 512×512 resolution, SiT-ATConv-XL/2 achieves better FID (7.88 vs 8.87) with 22.6% latency reduction compared to the vanilla SiT-XL, while improving both precision (0.69 vs 0.65) and recall (0.61 vs 0.56). Even remarkably, on ImageNet-1K with the REPA training technique, our full ATConv variants consistently outperform attention-based baselines: SiT-ATConv-XL/2 achieves significantly better FID (1.82 vs 1.97), higher IS (291.17 vs 282.33), and superior precision-recall trade-offs, all while reducing latency by 19%. These improvements scale consistently across all model sizes, with SiT-ATConv-B/2 reducing FID from 8.01 to 7.15 on ImageNet while accelerating inference by 18%. Besides, we show that by replacing only half of the attention with ATConv, the SiT-Hybrid variants can also produce remarkable improvements.

The consistent superiority of ATConv reveals a fundamental alignment with the denoising objective. ATConv constructs content-adaptive, *signed* spatial kernels with DKM that enforces mean-shifted filtering, naturally suppressing uninformative DC components while enhancing contrastive cues critical for ϵ -prediction. The signed kernels enable simultaneous

excitatory and inhibitory responses that preserve fine details during iterative denoising, while DKM’s competition dynamics stabilize gradients across varying noise levels. This inductive bias particularly benefits early denoising steps, explaining the observed superior sample quality.

Fig. 5 visually confirms these quantitative gains, with SiT-ATConv-XL/2 generating sharp facial details and coherent natural images despite using a purely convolutional architecture. These results suggest that the ATConv can serve as a new foundational operator for diffusion-based visual generation with both higher quality and efficiency. Fig. 6 further shows that the speed advantage of SiT-ATConv-XL to SiT-XL grows with increasing latent resolution. We hope that the excellent performance of ATConv can inspire the development of future generative architectures with higher efficiency and quality.

H. Ablation Studies

We conduct ablation studies to thoroughly examine the effectiveness of ATConv designs and the influence of different hyperparameter settings. All experiments are performed on the ImageNet-1K dataset following the protocol described in Sec. IV-C, with AttNet-T2 (27M parameters) chosen as the baseline. We report results in the following three aspects.

Effect of kernel size. Tab. IX-(a) investigates the impact of varying kernel sizes in ATConv. When using a uniform kernel size across all stages, the 3×3 configuration emerges as the most cost-effective, offering the best balance between accuracy and speed. Larger kernels (5×5 and 7×7) yield marginal accuracy gains (0.04% and 0.11%) but at a substantial cost in efficiency. Hierarchical configurations further reveal that simply enlarging receptive fields does not guaranty better performance. For instance, setting $[7, 7, 5, 3]$ across the four stages achieves the highest accuracy (84.53%), slightly outperforming the uniform 7×7 setting. This highlights that tailoring kernels to the intrinsic spatial properties of each stage is more effective than blindly enlarging them.

From another perspective, the accuracy improvements from larger ATConv kernels are modest compared to the drastic gains typically observed when scaling DWConv kernels. This suggests that ATConv, through its efficient integration of adaptive routing and lateral inhibition, already extracts rich visual representations using compact 3×3 kernels. In effect, ATConv breaks the traditional paradigm of pursuing ever-larger receptive fields and instead demonstrates a principled, efficient approach to spatial modeling.

Replace ATConv with alternative operators. Tab. IX-(b) compares ATConv with its alternative operators by replacing ATConv in AttNet-T2 with conventional convolutions or attention mechanisms. Replacing ATConv with vanilla Conv or DWConv (of various kernel sizes) leads to clear drops in Top-1 accuracy, underscoring ATConv’s superior accuracy–efficiency trade-off and its stronger representational capacity.

We also benchmark three leading attention mechanisms with linear spatial complexity: Hydra Attention, InLine Attention, and RankAug Attention. As shown in Tab. IX-(b), all three underperform ATConv in both accuracy and efficiency. For example, InLine and RankAug attention trail ATConv by

TABLE IX
QUANTITATIVE RESULTS OF ABLATION STUDIES.

(a) Ablation on Kernel Size of ATConv				
Kernel Size	Params (M)	FLOPs (G)	Thp. (fps)	Top-1 (%)
Unitary 3×3 (default)	27.01	5.11	1128	84.41
Unitary 5×5	27.02	5.15	953	84.45
Unitary 7×7	27.07	5.20	831	84.52
Hierarchical $[7, 5, 3, 3]$	27.02	5.14	947	84.46
Hierarchical $[7, 5, 5, 3]$	27.03	5.16	893	84.51
Hierarchical $[7, 7, 5, 3]$	27.03	5.17	865	84.53
(b) Ablation on Different Token Mixers				
Operator	Params (M)	FLOPs (G)	Thp. (fps)	Top-1 (%)
Default (3×3 ATConv)	27.01	5.11	1128	84.41
→ 3×3 Conv	40.04	7.35	787	81.81
→ 3×3 DWConv	20.52	3.98	1539	78.06
→ 5×5 DWConv	20.63	4.01	1311	79.11
→ 7×7 DWConv	20.79	4.06	1168	80.31
→ Hydra Attention [80]	29.18	5.47	772	79.83
→ InLine Attention [31]	31.63	5.63	684	83.61
→ RankAug Attention [81]	31.54	6.11	543	83.67
(c) Ablation on Building ATConv from DWConv				
Operator Config.	Params (M)	FLOPs (G)	Thp. (fps)	Top-1 (%)
3×3 DWConv	20.52	3.98	1539	78.06
+ Kernel Generator	22.64	4.37	1457	80.94
+ Last Linear Proj.	24.82	4.74	1338	81.65
+ Value Proj.	27.00	5.11	1190	83.17
+ Softmax on K	27.00	5.11	1139	—
→ Kernel Diff. on K	27.00	5.11	1168	82.80
→ Diff. Modulation on K	27.01	5.11	1128	84.41

0.80% and 0.74% in accuracy, while running $1.64\times$ and $2.0\times$ slower, respectively. These results confirm that by uniting the adaptivity of attention with the inductive bias of convolution, ATConv surpasses both traditional convolutions and advanced attention-based token mixers.

Roadmap from DWConv to ATConv. Finally, Tab. IX-(c) illustrates the progressive transformation from a standard 3×3 DWConv into ATConv, with performance gains measured at each step. Introducing the kernel generator, which converts static kernels into dynamic ones, yields a large accuracy boost (+2.88%) with negligible overhead. Adding a value projection and a final linear projection, thereby enabling adaptive routing coupled with dynamic kernels—further improves accuracy by 0.71% and 1.52%, respectively.

To validate the effectiveness of differential kernel modulation mechanism towards injecting the lateral inhibition attribute into convolutional calculation, we compare three alternatives: applying softmax on kernels (which causes training collapse), using a classic central difference [82] (which degrades performance due to indiscriminate suppression of low-frequency signals), and our differential kernel modulation. As reported in Tab. IX-(c), the latter delivers a further accuracy gain of +1.24% with minimal computational cost. These results demonstrate that each component of ATConv contributes meaningfully and consistently, collectively forging a highly effective visual operator.

V. CONCLUSION

This paper presents the first systematic identification of adaptive routing and lateral inhibition as the essential principles driving the success of attention mechanisms. Leveraging these insights, we introduce Attentive Convolution (ATConv), a purely convolutional operator that inherits these properties while preserving the efficiency and visual inductive biases of convolution. Our experiments show that ATConv not only surpasses leading attention mechanisms and Conv-attention hybrids in both accuracy and efficiency, but also establishes a scalable foundation for future architectures. We hope that this work will pave a new path for the evolution of convolutional architectures, bridging the gap with attention and inspiring future research in efficient visual modeling. We acknowledge two primary limitations of our current work. First, ATConv has not been explored in autoregressive settings, which constrains its applicability to next-token prediction paradigms prevalent in large language models. Second, while ATConv demonstrates computational advantages over most existing operators, its adaptive design introduces overhead compared to vanilla convolution. Future work will focus on extending ATConv to autoregressive frameworks and optimizing its efficiency through custom CUDA implementations.

REFERENCES

- [1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [2] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
- [3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.
- [4] Y. LeCun and Y. Bengio, "Convolutional networks for images, speech, and time series," *The handbook of brain theory and neural networks*, 1998.
- [5] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.
- [6] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.
- [7] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.
- [8] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *International conference on machine learning*. PMLR, 2021, pp. 10 347–10 357.
- [9] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.
- [10] A. Ali, H. Touvron, M. Caron, P. Bojanowski, M. Douze, A. Joulin, I. Laptev, N. Neverova, G. Synnaeve, J. Verbeek *et al.*, "Xcit: Cross-covariance image transformers," *Advances in neural information processing systems*, vol. 34, pp. 20 014–20 027, 2021.
- [11] X. Dong, J. Bao, D. Chen, W. Zhang, N. Yu, L. Yuan, D. Chen, and B. Guo, "Cswin transformer: A general vision transformer backbone with cross-shaped windows," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 12 124–12 134.
- [12] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 568–578.
- [13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.
- [14] P. K. A. Vasu, J. Gabriel, J. Zhu, O. Tuzel, and A. Ranjan, "Fastvit: A fast hybrid vision transformer using structural reparameterization," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 5785–5795.
- [15] A. Shaker, M. Maaz, H. Rasheed, S. Khan, M.-H. Yang, and F. S. Khan, "Swiftformer: Efficient additive attention for transformer-based real-time mobile vision applications," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 17 425–17 436.
- [16] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pvt v2: Improved baselines with pyramid vision transformer," *Computational visual media*, vol. 8, no. 3, pp. 415–424, 2022.
- [17] D. Shi, "Transnext: Robust foveal visual perception for vision transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 17 773–17 783.
- [18] Y. Wu, Y. Liu, X. Zhan, and M. Cheng, "P2t: Pyramid pooling transformer for scene understanding," *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, no. 11, pp. 12 760–12 771, 2022.
- [19] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1251–1258.
- [20] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.
- [21] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 11 976–11 986.
- [22] S. Woo, S. Debnath, R. Hu, X. Chen, Z. Liu, I. S. Kweon, and S. Xie, "Convnext v2: Co-designing and scaling convnets with masked autoencoders," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 16 133–16 142.
- [23] X. Ding, X. Zhang, N. Ma, J. Han, G. Ding, and J. Sun, "Repvgg: Making vgg-style convnets great again," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 13 733–13 742.
- [24] X. Ding, X. Zhang, J. Han, and G. Ding, "Scaling up your kernels to 31x31: Revisiting large kernel design in cnns," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 11 963–11 975.
- [25] C. Blakemore, R. H. Carpenter, and M. A. Georgeson, "Lateral inhibition between orientation detectors in the human visual system," *Nature*, vol. 228, no. 5266, pp. 37–39, 1970.
- [26] R. B. Tootell, N. K. Hadjikhani, W. Vanduffel, A. K. Liu, J. D. Mendola, M. I. Sereno, and A. M. Dale, "Functional analysis of primary visual cortex (v1) in humans," *Proceedings of the National Academy of Sciences*, vol. 95, no. 3, pp. 811–817, 1998.
- [27] T. Ojala, M. Pietikainen, and T. Maenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 24, no. 7, pp. 971–987, 2002.
- [28] G. Zhao and M. Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE transactions on pattern analysis and machine intelligence*, vol. 29, no. 6, pp. 915–928, 2007.
- [29] D. Marr and E. Hildreth, "Theory of edge detection," *Proceedings of the Royal Society of London. Series B. Biological Sciences*, vol. 207, no. 1167, pp. 187–217, 1980.
- [30] Z. Tu, H. Talebi, H. Zhang, F. Yang, P. Milanfar, A. Bovik, and Y. Li, "Maxvit: Multi-axis vision transformer," in *European conference on computer vision*. Springer, 2022, pp. 459–479.
- [31] D. Han, Y. Pu, Z. Xia, Y. Han, X. Pan, X. Li, J. Lu, S. Song, and G. Huang, "Bridging the divide: Reconsidering softmax and linear attention," *Advances in Neural Information Processing Systems*, vol. 37, pp. 79 221–79 245, 2024.
- [32] Z. Qin, W. Sun, H. Deng, D. Li, Y. Wei, B. Lv, J. Yan, L. Kong, and Y. Zhong, "cosformer: Rethinking softmax in attention," *arXiv preprint arXiv:2202.08791*, 2022.
- [33] P. Shaw, J. Uszkoreit, and A. Vaswani, "Self-attention with relative position representations," *arXiv preprint arXiv:1803.02155*, 2018.
- [34] B. Heo, S. Park, D. Han, and S. Yun, "Rotary position embedding for vision transformer," in *European Conference on Computer Vision*. Springer, 2024, pp. 289–305.
- [35] Z. Liu, H. Hu, Y. Lin, Z. Yao, Z. Xie, Y. Wei, J. Ning, Y. Cao, Z. Zhang, L. Dong *et al.*, "Swin transformer v2: Scaling up capacity and

- resolution,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 12 009–12 019.
- [36] B. Graham, A. El-Nouby, H. Touvron, P. Stock, A. Joulin, H. Jégou, and M. Douze, “Levit: a vision transformer in convnet’s clothing for faster inference,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 12 259–12 269.
- [37] Z. Dai, H. Liu, Q. V. Le, and M. Tan, “Coatnet: Marrying convolution and attention for all data sizes,” *Advances in neural information processing systems*, vol. 34, pp. 3965–3977, 2021.
- [38] H. Wu, B. Xiao, N. Codella, M. Liu, X. Dai, L. Yuan, and L. Zhang, “Cvt: Introducing convolutions to vision transformers,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 22–31.
- [39] Y. LeCun, B. Boser, J. Denker, D. Henderson, R. Howard, W. Hubbard, and L. Jackel, “Handwritten digit recognition with a back-propagation network,” *Advances in neural information processing systems*, vol. 2, 1989.
- [40] M. Tan and Q. Le, “Efficientnet: Rethinking model scaling for convolutional neural networks,” in *International conference on machine learning*. PMLR, 2019, pp. 6105–6114.
- [41] W. Yu, P. Zhou, S. Yan, and X. Wang, “Inceptionnext: When inception meets convnext,” in *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, 2024, pp. 5672–5683.
- [42] H. Yu, H. Chen, W. Peng, X. Cheng, and G. Zhao, “Freenet: Liberating depth-wise separable operations for building faster mobile vision architectures,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 9, 2025, pp. 9607–9615.
- [43] B. Yang, G. Bender, Q. V. Le, and J. Ngiam, “Concconv: Conditionally parameterized convolutions for efficient inference,” *Advances in neural information processing systems*, vol. 32, 2019.
- [44] Y. Chen, X. Dai, M. Liu, D. Chen, L. Yuan, and Z. Liu, “Dynamic convolution: Attention over convolution kernels,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 030–11 039.
- [45] N. Ma, X. Zhang, J. Huang, and J. Sun, “Weightnet: Revisiting the design space of weight networks,” in *European conference on computer vision*. Springer, 2020, pp. 776–792.
- [46] C. Li, A. Zhou, and A. Yao, “Omni-dimensional dynamic convolution,” *arXiv preprint arXiv:2209.07947*, 2022.
- [47] D. Li, J. Hu, C. Wang, X. Li, Q. She, L. Zhu, T. Zhang, and Q. Chen, “Involution: Inverting the inheritance of convolution for visual recognition,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 12 321–12 330.
- [48] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.
- [49] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, “Cbam: Convolutional block attention module,” in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.
- [50] W. Hua, Z. Dai, H. Liu, and Q. Le, “Transformer quality in linear time,” in *International conference on machine learning*. PMLR, 2022, pp. 9099–9117.
- [51] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar *et al.*, “Llama: Open and efficient foundation language models,” *arXiv preprint arXiv:2302.13971*, 2023.
- [52] R. Anil, A. M. Dai, O. Firat, M. Johnson, D. Lepikhin, A. Passos, S. Shakeri, E. Taropa, P. Bailey, Z. Chen *et al.*, “Palm 2 technical report,” *arXiv preprint arXiv:2305.10403*, 2023.
- [53] L. Zhu, X. Wang, Z. Ke, W. Zhang, and R. W. Lau, “Biformer: Vision transformer with bi-level routing attention,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 10 323–10 333.
- [54] Y. Li, J. Hu, Y. Wen, G. Evangelidis, K. Salahi, Y. Wang, S. Tulyakov, and J. Ren, “Rethinking vision transformers for mobilenet size and speed,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 16 889–16 900.
- [55] J. Yang, C. Li, P. Zhang, X. Dai, B. Xiao, L. Yuan, and J. Gao, “Focal self-attention for local-global interactions in vision transformers,” *arXiv preprint arXiv:2107.00641*, 2021.
- [56] H. Huang, X. Zhou, J. Cao, R. He, and T. Tan, “Vision transformer with super token sampling,” *arXiv preprint arXiv:2211.11167*, 2022.
- [57] J. Guo, K. Han, H. Tang, X. Chen, Y. Wang, and C. Xu, “Cmt: Convolutional neural networks meet vision transformers,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 12 175–12 185.
- [58] K. Li, Y. Wang, J. Zhang, P. Gao, G. Song, Y. Liu, H. Li, and Y. Qiao, “Uniformer: Unifying convolution and self-attention for visual recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 10, pp. 12 581–12 600, 2023.
- [59] Z. Xia, X. Pan, S. Song, L. E. Li, and G. Huang, “Vision transformer with deformable attention,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 4794–4803.
- [60] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.
- [61] I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” *arXiv preprint arXiv:1711.05101*, 2017.
- [62] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, “Microsoft coco: Common objects in context,” in *European conference on computer vision*. Springer, 2014, pp. 740–755.
- [63] K. He, G. Gkioxari, P. Dollár, and R. Girshick, “Mask r-cnn,” in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.
- [64] Z. Cai and N. Vasconcelos, “Cascade r-cnn: High quality object detection and instance segmentation,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 5, pp. 1483–1498, 2019.
- [65] Y. Zhang and H. Wang, “Diverse embedding expansion network and low-light cross-modality benchmark for visible-infrared person re-identification,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 2153–2162.
- [66] X. Lin, J. Li, Z. Ma, H. Li, S. Li, K. Xu, G. Lu, and D. Zhang, “Learning modal-invariant and temporal-memory for video-based visible-infrared person re-identification,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 20 973–20 982.
- [67] T. Karras, S. Laine, and T. Aila, “A style-based generator architecture for generative adversarial networks,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4401–4410.
- [68] S. Yu, S. Kwak, H. Jang, J. Jeong, J. Huang, J. Shin, and S. Xie, “Representation alignment for generation: Training diffusion transformers is easier than you think,” *arXiv preprint arXiv:2410.06940*, 2024.
- [69] W. Peebles and S. Xie, “Scalable diffusion models with transformers,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 4195–4205.
- [70] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, “Scene parsing through ade20k dataset,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 633–641.
- [71] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. Hoi, “Deep learning for person re-identification: A survey and outlook,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 6, pp. 2872–2893, 2021.
- [72] M. Ye, W. Ruan, B. Du, and M. Z. Shou, “Channel augmented joint learning for visible-infrared recognition,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 13 567–13 576.
- [73] H. Yu, X. Cheng, W. Peng, W. Liu, and G. Zhao, “Modality unifying network for visible-infrared person re-identification,” in *ICCV*, 2023, pp. 11 185–11 195.
- [74] Y. Jiang, X. Cheng, H. Yu, X. Liu, H. Chen, and G. Zhao, “Domain shifting: A generalized solution for heterogeneous cross-modality person re-identification,” in *European Conference on Computer Vision*. Springer, 2025, pp. 289–306.
- [75] Y. Jiang, H. Yu, X. Cheng, H. Chen, Z. Sun, and G. Zhao, “From laboratory to real world: A new benchmark towards privacy-preserved visible-infrared person re-identification,” in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 8828–8837.
- [76] F.-A. Croitoru, V. Hondru, R. T. Ionescu, and M. Shah, “Diffusion models in vision: A survey,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, no. 9, pp. 10 850–10 869, 2023.
- [77] C. He, Y. Shen, C. Fang, F. Xiao, L. Tang, Y. Zhang, W. Zuo, Z. Guo, and X. Li, “Diffusion models in low-level vision: A survey,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.
- [78] Z. Cao, F. Hong, T. Wu, L. Pan, and Z. Liu, “Diff++: 3d-aware diffusion transformer for large-vocabulary 3d generation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.
- [79] M. Li, Y. Fu, T. Zhang, J. Liu, D. Dou, C. Yan, and Y. Zhang, “Latent diffusion enhanced rectangle transformer for hyperspectral image restoration,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

- [80] D. Bolya, C.-Y. Fu, X. Dai, P. Zhang, and J. Hoffman, “Hydra attention: Efficient attention with many heads,” in *European conference on computer vision*. Springer, 2022, pp. 35–49.
- [81] Q. Fan, H. Huang, and R. He, “Breaking the low-rank dilemma of linear attention,” in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 25 271–25 280.
- [82] Z. Su, W. Liu, Z. Yu, D. Hu, Q. Liao, Q. Tian, M. Pietikäinen, and L. Liu, “Pixel difference networks for efficient edge detection,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 5117–5127.