TransNeXt: Robust Foveal Visual Perception for Vision Transformers

Dai Shi
daishiresearch@gmail.com
Code: https://github.com/DaiShiResearch/TransNeXt

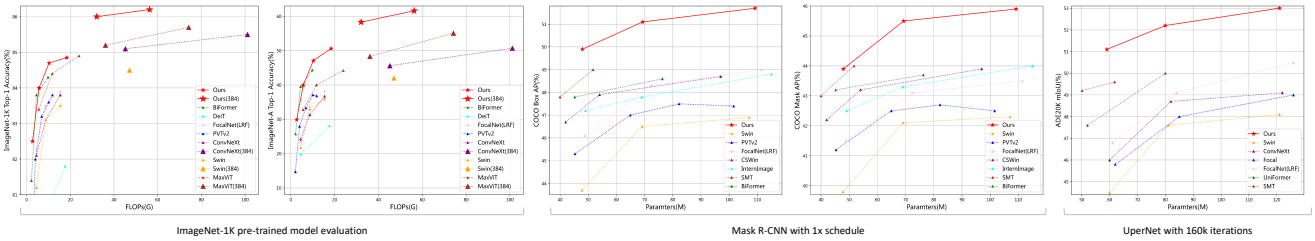

Figure 1. A comprehensive comparison of performance on ImageNet-1K, robustness on ImageNet-A, COCO detection and instance segmentation performance based on Mask R-CNN 1×, ADE20K semantic segmentation performance based on UperNet.

Abstract

Due to the depth degradation effect in residual connections, many efficient Vision Transformers models that rely on stacking layers for information exchange often fail to form sufficient information mixing, leading to unnatural visual perception. To address this issue, in this paper, we propose ***Aggregated Attention****, a biomimetic design-based token mixer that simulates biological foveal vision and continuous eye movement while enabling each token on the feature map to have a global perception. Furthermore, we incorporate learnable tokens that interact with conventional queries and keys, which further diversifies the generation of affinity matrices beyond merely relying on the similarity between queries and keys. Our approach does not rely on stacking for information exchange, thus effectively avoiding depth degradation and achieving natural visual perception. Additionally, we propose* ***Convolutional GLU****, a channel mixer that bridges the gap between GLU and SE mechanism, which empowers each token to have channel attention based on its nearest neighbor image features, enhancing local modeling capability and model robustness. We combine aggregated attention and convolutional GLU to create a new visual backbone called* ***TransNeXt****. Extensive experiments demonstrate that our TransNeXt achieves state-of-the-art performance across multiple model sizes. At a resolution of* 224^2*, TransNeXt-Tiny attains an ImageNet accuracy of* ***84.0%****, surpassing ConvNeXt-B with* ***69%*** *fewer parameters. Our TransNeXt-Base achieves an ImageNet accuracy of* ***86.2%*** *and an ImageNet-A accuracy of* ***61.6%*** *at a resolution of* 384^2*, a COCO object detection mAP of* ***57.1****, and an ADE20K semantic segmentation mIoU of* ***54.7****.*

1. Introduction

The Vision Transformer (ViT) [12] has emerged as a popular backbone architecture for various computer vision tasks in recent years. The ViT model comprises two key components: the self-attention layer (token mixer) and the MLP layer (channel mixer). The self-attention mechanism plays a crucial role in feature extraction by dynamically generating an affinity matrix through similarity computations between queries and keys. This global information aggregation method has demonstrated remarkable feature extraction potential, with no inductive bias like convolution [25], and can build powerful data-driven models. However, the transformer encoder design of vision transformers, originally developed for language modeling [44], exhibits inherent limitations in downstream computer vision tasks. Specifically, the computation of the global affinity matrix in self-attention poses a challenge due to its quadratic complexity and high memory consumption, which restricts its application on high-resolution image features.

In order to mitigate the computational and memory burdens imposed by the quadratic complexity inherent in the self-attention mechanism, a plethora of sparse attention mechanisms have been proposed in previous studies. One such representative method is local attention [33], which restricts attention within a window on the feature map. However, due to the limited receptive field, this method often requires alternating stacking with different types of token mixers to achieve cross-window information exchange. Another representative method spatially downsamples the keys and values of attention (such as pooling [47–49], grid sampling [43]). This method, due to its sacrifice of the query's fine-grained perception of the feature map, also has certain

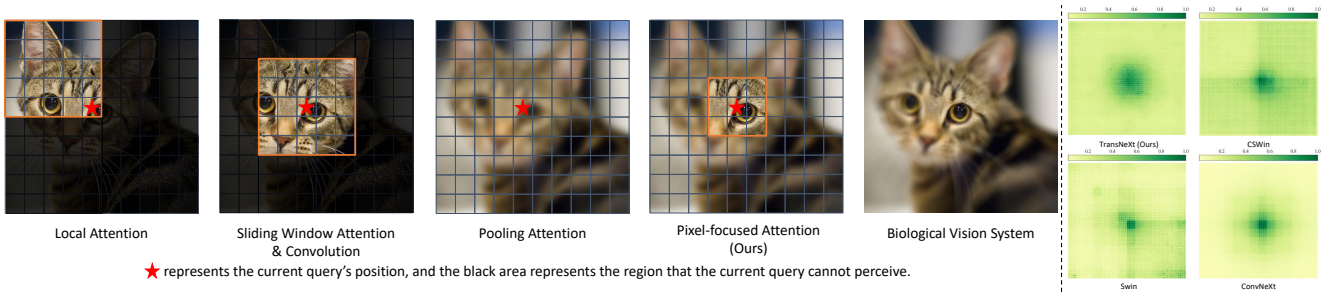

Figure 2. A comparison of prevalent visual information aggregation mechanisms, our proposed method, and biological visual systems (Left) and a visualization comparison of the Effective Receptive Field [35] between our method and the prevalent backbone networks, using the output at stage 3 (Right). Each ERF image is generated by averaging over 5000 224^2-sized images from ImageNet-1K validation set.

limitations. Recent studies [5, 43] have alternately stacked spatial downsampling attention and local attention, achieving commendable performance results.

However, recent studies [8, 45] and experiments [22] have shown that deep networks with residual blocks [14] behave like ensembles of shallower networks, indicating that the cross-layer information exchange achieved by stacking blocks may not be as effective as anticipated.

On the other hand, both local attention and spatial downsampling attention differ significantly from the workings of biological vision. Biological vision possesses higher acuity for features around the visual focus and lower acuity for distant features. Moreover, as the eyeball moves, this characteristic of biological vision remains consistent for pixels at any position in the image, implying pixel-wise translational equivariance. However, in local attention based on window partitioning, tokens at the window edge and center are not treated equivalently, presenting a clear discrepancy.

We have observed that due to depth degradation effects, many efficient ViT models are unable to form sufficient information mixing through stacking. Even with a deep stack of layers, the traces of their window partitioning always form unnatural artifacts, as shown in Fig 2. To address this issue, we investigate a visual modeling approach that closely aligns with biological vision to mitigates potential model depth degradation and achieve information perception closer to human foveal vision. To this end, we initially introduce **Pixel-focused Attention**, which employs a dual-path design. In one path, each query has fine-grained attention to its nearest neighbor features, while in the other path, each query has coarse-grained attention to spatial downsampled features, allowing for a global perception. This approach operates on a per-pixel basis, effectively simulating the continuous movement of the eyeball. Furthermore, we incorporate query embedding and positional attention mechanisms into pixel-focused attention, leading to the proposal of **Aggregated Pixel-focused Attention**, which we abbreviate as **Aggregated Attention**. This approach further diversifies the generation of affinity matrices beyond merely relying on the similarity between queries and keys, thereby achieving the aggregation of multiple attention mechanisms within a single attention layer. We also reevaluate the design requirements of the channel mixer in vision transformers and propose a novel channel mixer named **Convolutional GLU**. This mixer is more apt for image tasks and integrates local feature-based channel attention to enhance model robustness.

We introduce **TransNeXt**, a hierarchical visual backbone network that incorporates **aggregated attention** as a token mixer and **convolutional GLU** as a channel mixer. Through comprehensive evaluation across image classification, object detection, and segmentation tasks, we demonstrate the efficacy of these mixing components. Our TransNeXt-Tiny, pretrained solely on ImageNet-1K, achieves an ImageNet accuracy of **84.0%**, surpassing ConvNeXt-B. In COCO object detection, it attains a box mAP of **55.1** using the DINO detection head, outperforming ConvNeXt-L pretrained at a resolution of 384^2 by 1.7. Our TransNeXt-Small/Base, fine-tuned at a resolution of 384^2 for merely **5 epochs**, achieves an ImageNet accuracy of **86.0%/86.2%**, surpassing the previous state-of-the-art MaxViT-Base fine-tuned for 30 epochs by 0.3%/0.5%. Moreover, when evaluated on the highly challenging ImageNet-A test set at a resolution of 384^2, our TransNeXt-Small/Base models achieve an impressive top-1 accuracy of **58.3%/61.6%**, significantly outperforming ConvNeXt-L by 7.6%/10.9%, setting a new benchmark of robustness for ImageNet-1K supervised models.

In summary, our contributions are as follows:

1. Proposing **pixel-focused attention**, a token mixer closely aligns with biological foveal vision and mitigates potential model depth degradation. This novel attention mechanism works on a per-pixel basis, effectively simulating the continuous movement of the eyeball and highly aligning with the focal perception mode of biological vision. It possesses visual priors comparable to convolution.
2. Proposing **aggregated attention**, an enhanced version of pixel-focused attention, which further aggregates two types of non-QKV attention mechanisms into pixel-focused attention. Notably, we propose a highly efficient

approach within this framework, with the additional computational overhead accounting for a mere 0.2%-0.3% of the entire model, leading to a exceptionally cost-effective unification of QKV attention, LKV attention, and QLV attention within a single mixer layer.

3. Proposing **length-scaled cosine attention** that enhances the extrapolation capability of existing attention mechanisms for multi-scale input. This allows TransNeXt to achieve superior large-scale image extrapolation performance compared to pure convolutional networks.
4. Proposing **convolutional GLU**, which incorporates channel attention based on nearest neighbor image features. In comparison to convolutional feed-forward, it realizes the attentionalization of the channel mixer with fewer FLOPs, thereby effectively enhancing the model's robustness.
5. Introducing **TransNeXt**, a visual backbone that delivers state-of-the-art performance in various visual tasks such as image classification, object detection, and semantic segmentation among models of similar size. It also exhibits state-of-the-art robustness.

2. Related Work

Vision transformers: Vision Transformer (ViT) [12] was the first to introduce transformer architecture to visual tasks, where images are segmented into non-overlapping patches and subsequently linearly projected into token sequences, which are later encoded by a transformer encoder. When trained with large-scale pretraining data or thoughtfully designed training strategies, ViT models outperform convolutional neural networks (CNNs)[14, 24, 25], exhibiting remarkable performance in image classification and other downstream tasks.

Non-QKV attention variants: In self-attention, the dynamic affinity matrix is generated through the interaction between queries and keys. Recently, several studies [1, 26, 41, 53] have explored the use of learnable tokens as a replacement for the original queries or keys to generate dynamic affinity matrices. Involution [26] and VOLO [53], for instance, use learnable tokens to replace the original keys, resulting in dynamic affinity matrices that are exclusively correlated with queries. In contrast, QnA [1] utilizes learnable tokens to replace queries, leading to dynamic affinity matrices that are only correlated with keys. Both methods have shown effectiveness.

Biomimetic vision modeling: Human vision exhibits higher acuity for features around the visual focus and lower acuity for distant features. This biomimetic design has been integrated into several machine vision models [36, 51, 52]. Specifically, Focal Transformer [51] designs a visual attention based on this concept, but it operates based on window partitioning. Tokens located at the window edges cannot obtain natural foveal vision, and its window-wise manner cannot simulate the continuous movement of the human eyeball. Our approach effectively addresses these shortcomings.

3. Method

3.1. Aggregated Pixel-focused Attention

3.1.1 Pixel-focused Attention

Inspired by the functioning of biological visual systems, we have designed a pixel-focused attention mechanism that possesses fine-grained perception in the vicinity of each query, while concurrently maintaining a coarse-grained awareness of global information. To achieve the pixel-wise translational equivariance inherent in eyeball movements, we employ a dual-path design incorporating query-centered sliding window attention and pooling attention. Furthermore, to induce coupling between the two attention paths, we compute the importance in the same softmax for the query-key similarity results of both paths. This results in a competition between fine-grained and coarse-grained features, transforming pixel-focused attention into a multi-scale attention mechanism.

Given an input $X \in \mathbb{R}^{C\times H\times W}$, we now focus on the operations performed on a single pixel in the input feature map. We define a set of pixels within a sliding window centered at pixel at (i, j) as $\rho(i, j)$. For a fixed window size of $k \times k$, $\|\rho(i, j)\| = k^2$. Concurrently, we define the set of pixels obtained from pooling the feature map as $\sigma(X)$. Given a pooling size of $H_p \times W_p$, $\|\sigma(X)\| = H_pW_p$. Therefore, **p**ixel-**f**ocused **a**ttention (**PFA**) can be described as follows:

$$\begin{aligned} S_{(i,j)\sim\rho(i,j)} &= Q_{(i,j)}K^T_{\rho(i,j)} \\ S_{(i,j)\sim\sigma(X)} &= Q_{(i,j)}K^T_{\sigma(X)} \end{aligned} \tag{1}$$

$$\begin{aligned} A_{(i,j)} =& \text{softmax} \\ & (\frac{\text{Concat}(S_{(i,j)\sim\rho(i,j)}, S_{(i,j)\sim\sigma(X)})}{\sqrt{d}} + B_{(i,j)}) \end{aligned} \tag{2}$$

$$\begin{aligned} A_{(i,j)\sim\rho(i,j)}, A_{(i,j)\sim\sigma(X)} =& \text{Split}(A_{(i,j)}) \\ & \text{with size } [k^2, H_pW_p] \end{aligned} \tag{3}$$

$$\textbf{PFA}(X_{(i,j)}) = A_{(i,j)\sim\rho(i,j)}V_{\rho(i,j)} + A_{(i,j)\sim\sigma(X)}V_{\sigma(X)} \tag{4}$$

Activate and Pool: In order to utilize the linear complexity mode of PFA for large-scale image inference in subsequent applications, we employ parameter-free adaptive average pooling for downsampling in the spatial dimension. However, the average pooling operator significantly loses information. Therefore, we use a single-layer neural network for projection and activation before feature map pooling to compress and extract useful information in advance, thereby improving the information compression rate after downsampling. After pooling, we once again use layer normalization to normalize the output to ensure the variance consistency of X and $\sigma(X)$. The downsampling operator we propose,

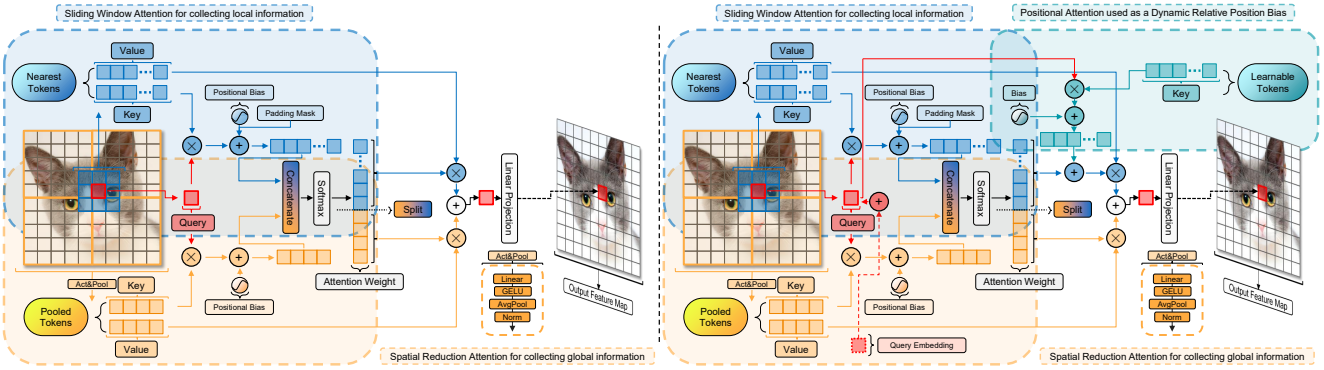

Figure 3. An illustration of the comparison between pixel-focused attention (left) and aggregated attention (right). Both have a feature size of 10×10, a window size of 3×3, and a pool size of 2×2.

termed 'Activate and Pool', can be expressed by the following equation:

$$\sigma(X) = \text{LayerNorm}(\text{AvgPool}(\text{GELU}(\text{Linear}(X)))) \tag{5}$$

We replaced the downsampling module in PVTv2-li [48] with our 'activate and pool' mechanism and designed a 2M-sized model for ablation experiments on CIFAR-100 [23]. Our module improved the top-1 accuracy of PVTv2-li from 68.1% to 70.4%, demonstrating the effectiveness of this approach.

Padding mask: In the sliding window path, pixels located at the edge of the feature map inevitably compute similarities with zero-padding outside the boundary. To prevent these zero similarities from influencing the softmax operation, we employ a padding mask to set these results to $-\infty$.

3.1.2 Aggregating Diverse Attentions in a Single Mixer

Query embedding: Several vision-language models [27, 28] utilize queries originating from the textual modality to perform cross-attention on keys derived from the visual modality, thereby achieving cross-modal information aggregation to complete Visual Question Answering (VQA) tasks. Moreover, it has been proven effective and efficient to incorporate and optimize learnable prefix query tokens when fine-tuning these multimodal models to adapt to specific subtasks.

A natural extension of this idea is to incorporate these learnable query tokens into the attention mechanism of the backbone network for well-defined tasks such as image classification, object detection, and semantic segmentation, and directly optimize them. This approach has been validated by previous work [1] for its effectiveness.

This method differs from traditional QKV attention as it does not use queries from the input but learns a query defined by the current task to perform cross-attention. Therefore, we categorize this method as **L**earnable-**K**ey-**V**alue (**LKV**) attention, drawing a parallel to QKV attention. We found that adding a learnable **Q**uery **E**mbedding (**QE**) to all query tokens in traditional QKV attention can achieve similar information aggregation effects with negligible additional overhead. We only need to modify Equation 1 as follows:

$$\begin{aligned} S_{(i,j)\sim\rho(i,j)} &= (Q_{(i,j)} + \text{QE})K^T_{\rho(i,j)} \\ S_{(i,j)\sim\sigma(X)} &= (Q_{(i,j)} + \text{QE})K^T_{\sigma(X)} \end{aligned} \tag{6}$$

Positional attention: An alternative approach to information aggregation is the use of a set of learnable keys that interact with queries originating from the input to obtain attention weights, *i.e.*, **Q**uery-**L**earnable-**V**alue (**QLV**) attention. This method differs from traditional QKV attention as it disrupts the one-to-one correspondence between keys and values, resulting in learning more implicit relative positional information for the current query. Consequently, it is often employed in conjunction with a sliding window in visual tasks [26, 53]. Unlike static affinity matrices such as convolution or relative position bias, the affinity matrix generated in this way takes into account the impact of the current query and can dynamically adapt based on it. We have observed that this data-driven modeling approach exhibits greater robustness compared to static relative position bias and can further enhance locality modeling capabilities. Leveraging this feature, we introduce a set of learnable tokens $T \in \mathbb{R}^{d\times k^2}$ in each attention head, allowing these tokens to interact with queries to obtain additional dynamic position bias and add it to $A_{(i,j)\sim\rho(i,j)}$. Using this enhancement only requires an additional computational overhead of HWk^2C. We only need to modify Equation 4 as follows:

$$\begin{aligned} \textbf{PFA}(X_{(i,j)}) =& (A_{(i,j)\sim\rho(i,j)} + Q_{(i,j)}T)V_{\rho(i,j)} \\ &+ A_{(i,j)\sim\sigma(X)}V_{\sigma(X)} \end{aligned} \tag{7}$$

3.1.3 Overcoming Multi-scale Image Input

Length-scaled cosine attention: In contrast to the scaled dot product attention, the scaled cosine attention, which employs cosine similarity, has been observed to generate more

moderate attention weights [19, 33] and effectively enhance the training stability of large visual models [9]. The scaled cosine attention typically multiplies an additional learnable coefficient λ to the cosine similarity results of queries and keys, enabling the attention mechanism to effectively ignore insignificant tokens [19]. Recent studies [3, 13] have discovered that as the length of the input sequence increases, the confidence of the attention output decreases. Therefore, the scaling factor of the attention mechanism should be related to the length of the input sequence [3]. [40] further proposed that the design of attention should exhibit entropy invariance to facilitate better generalization to unknown lengths. [40] provided an estimate of the entropy of the scaled dot product attention with a sequence length n when queries and keys are approximated as vectors with a magnitude of $\sqrt{d}$:

$$\mathcal{H}_i \approx \log n - 0.24\lambda d + \mathcal{O}(1) \tag{8}$$

For cosine similarity, we define the queries and keys with ℓ_2-normalization applied along their head dimensions as $\hat{Q}$ and $\hat{K}$ respectively, both of which have magnitudes of 1. To maintain entropy invariance and disregard constant terms, we set $\lambda \approx \frac{\log n}{0.24}$. Given that Equation 8 is merely an estimate, we set $\lambda = \tau \log n$, where τ is a learnable variable initialized to $\frac{1}{0.24}$ for each attention head. We propose **length-scaled cosine attention** as follows:

$$\text{Attention}(Q, K, V) = \text{softmax}(\tau \log N * \hat{Q}\hat{K}^T)V \tag{9}$$

Here, N denotes the count of effective keys each query interacts with, excluding the count of masked tokens. Specifically, when applied in a transformer decoder [44], future tokens masked by a causal mask should not be counted in N. In the context of pixel-focused attention, N is calculated as $N_{(i,j)} = \|\rho(i,j)\| + \|\sigma(X)\| - \|\mu(i,j)\|$, where $\mu(i,j)$ represents the set of padding-masked tokens at position (i,j).

Position bias: To further enhance the extrapolation capability of pixel-focused attention for multi-scale image inputs, we employ different methods to calculate $B_{(i,j)\sim\rho(i,j)}$ and $B_{(i,j)\sim\sigma(X)}$ on two paths. On the pooling feature path, we use **log**-spaced **c**ontinuous **p**osition **b**ias (**log-CPB**) [33], a 2-layer MLP with a ReLU [37] to compute $B_{(i,j)\sim\sigma(X)}$ from the spatial relative coordinates $\Delta_{(i,j)\sim\sigma(X)}$ between $Q_{(i,j)}$ and $K_{\sigma(X)}$. On the sliding window path, we directly use a learnable $B_{(i,j)\sim\rho(i,j)}$. On one hand, this is because the size of the sliding window is fixed and does not require extrapolation of unknown relative position biases through log-CPB, thus saving computational resources. On the other hand, we observe that using log-CPB to calculate $B_{(i,j)\sim\rho(i,j)}$ results in performance degradation. We believe this is because $\Delta_{(i,j)\sim\sigma(X)}$ represents the spatial relative coordinates between fine-grained tokens and coarse-grained tokens, while $\Delta_{(i,j)\sim\rho(i,j)}$represents the spatial relative coordinates between fine-grained tokens, and their numerical meanings are different. We discuss these details further in appendix.

Aggregated attention: By applying the aforementioned diverse attention aggregation methods and techniques for enhancing the extrapolation capability for multi-scale inputs, we propose an enhanced version of pixel-focused attention, termed aggregated pixel-focused attention, which we abbreviate as **A**ggregated **A**ttention (**AA**). It can be described as follows:

$$\begin{aligned} S_{(i,j)\sim\rho(i,j)} &= (\hat{Q}_{(i,j)} + \text{QE})\hat{K}^T_{\rho(i,j)} \\ S_{(i,j)\sim\sigma(X)} &= (\hat{Q}_{(i,j)} + \text{QE})\hat{K}^T_{\sigma(X)} \end{aligned} \tag{10}$$

$$B_{(i,j)} = \text{Concat}(B_{(i,j)\sim\rho(i,j)}, \textbf{log-CPB}(\Delta_{(i,j)\sim\sigma(X)})) \tag{11}$$

$$\begin{aligned} A_{(i,j)} =& \text{softmax}(\tau \log N * \\ & \text{Concat}(S_{(i,j)\sim\rho(i,j)}, S_{(i,j)\sim\sigma(X)}) + B_{(i,j)}) \end{aligned} \tag{12}$$

$$\begin{aligned} A_{(i,j)\sim\rho(i,j)}, A_{(i,j)\sim\sigma(X)} =& \text{Split}(A_{(i,j)}) \\ & \text{with size } [k^2, H_pW_p] \end{aligned} \tag{13}$$

$$\begin{aligned} \textbf{AA}(X_{(i,j)}) =& (A_{(i,j)\sim\rho(i,j)} + \hat{Q}_{(i,j)}T)V_{\rho(i,j)} \\ & + A_{(i,j)\sim\sigma(X)}V_{\sigma(X)} \end{aligned} \tag{14}$$

3.1.4 Feature Analysis

Computational complexity: Given an input $X \in \mathbb{R}^{C\times H\times W}$, a pooling size of $H_p \times W_p$, and a window size of $k \times k$, we consider the impact of 'activate and pool' operation and linear projection. The computational complexities of pixel-focused attention and aggregated attention are:

$$\begin{aligned} \Omega(\textbf{PFA}) =& 5HWC^2 + 2H_pWpC^2 \\ & + 2HWH_pW_pC + 2HWk^2C \end{aligned} \tag{15}$$

$$\begin{aligned} \Omega(\textbf{AA}) &= \Omega(\textbf{PFA}) + HWk^2C \\ &= 5HWC^2 + 2H_pWpC^2 \\ &\quad + 2HWH_pW_pC + 3HWk^2C \end{aligned} \tag{16}$$

We observe that when the pooling size $H_p \times W_p$ is set to a value independent of the input size, Both $\Omega(\textbf{PFA})$ and $\Omega(\textbf{AA})$ scales linearly with the length of the input sequence. This implies that both PFA and AA can perform inference in a **linear complexity mode**.

Optimal accuracy-efficiency trade-off: Through empirical studies, we observed that the size of the sliding window has a negligible impact on model performance. Consequently, we employed the minimal form of a 3×3 sliding window to capture features near the visual focus, significantly reducing computational and memory consumption. We attribute this to the presence of pooling feature paths, which endow each query with a global receptive field, thereby greatly diminishing the need to expand the sliding window size to extend the receptive field. Detailed ablation study results and discussions can be found in appendix.

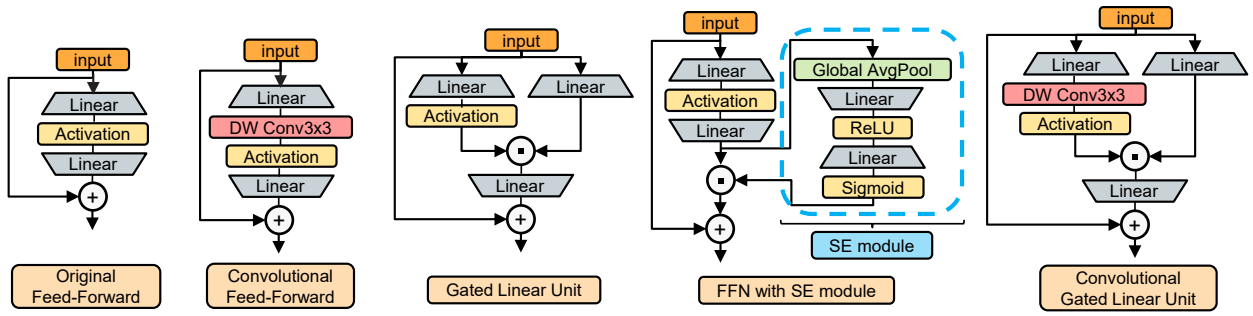

Figure 4. Comparison of prevalent channel mixer designs and Convolutional GLU

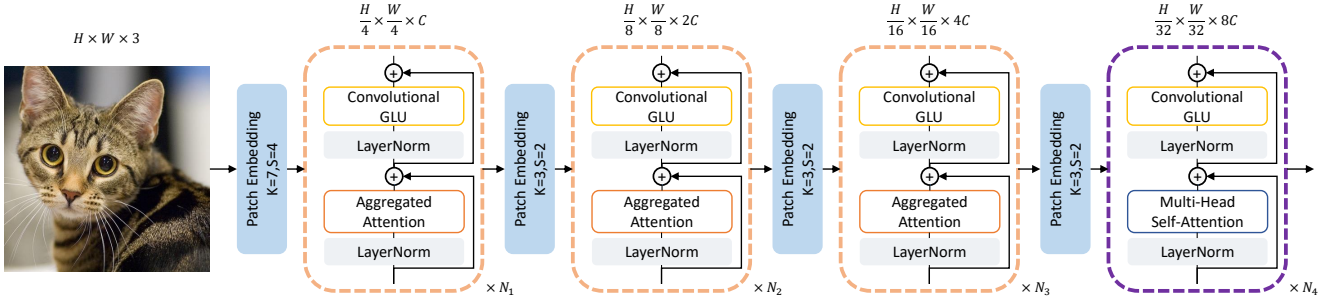

Figure 5. An illustration of TrasnNeXt architecture.

3.2. Convolutional GLU

3.2.1 Motivation

Gated channel attention in ViT era: Previous work, represented by the Squeeze-and-Excitation (SE) mechanism [20], first introduced channel attention into the field of computer vision, which uses a branch with an activation function to gate the network output. In gated channel attention, the gating branch has more decision-making power than the value branch, and it ultimately determines whether the corresponding output elements are zeroed. From this perspective, the SE mechanism cleverly uses features after global average pooling as the input of the gating branch, achieving a largest receptive field for better decision-making and solving the problem of insufficient receptive field in CNNs structures at the same time. However, in the ViT era, global receptive fields are no longer scarce. Various global token mixers represented by self-attention have achieved higher quality global information aggregation than global average pooling. This makes the global pooling method used by the SE mechanism show some shortcomings, such as this method makes all tokens on the feature map share the same gating signal, making its channel attention lack flexibility and too coarse-grained. Despite this, it's worth noting that ViT structures lack channel attention. Recent research [56] has found that incorporating the SE mechanism into a channel mixer can effectively enhance model robustness, as shown in Fig. 4.

Convolution in ViT era: Recent studies [6, 21] have shown that introducing a 3×3 depthwise convolution [4] into the vision transformer can be viewed as a form of conditional position encoding (CPE) [6], which effectively captures positional information from zero-padding.

3.2.2 Rethinking Channel Mixer Design

The Gated Linear Unit (GLU) [7, 39] is a channel mixer that has been shown to outperform Multi-Layer Perceptron (MLP) in various natural language processing tasks. GLU consists of two linear projections that are element-wise multiplied, with one projection being activated by a gating function. Unlike the SE mechanism, its gating signal for each token is derived from the token itself and does not have a larger receptive field than the value branch.

More elegant design: We found that simply adding a minimal form of 3×3 depthwise convolution before the activation function of GLU's gating branch can make its structure conform to the design concept of gated channel attention and convert it into a gated channel attention mechanism based on nearest neighbor features. We named this method **Convolutional GLU**, as shown in Fig. 4.

Feature analysis: Each token in **Conv**olutional **GLU** (**ConvGLU**) possesses a unique gating signal, based on its nearest fine-grained features. This addresses the overly coarse-grained drawback of the global average pooling in the SE mechanism. It also meets the needs of some ViT models without position encoding design that require position information provided by depthwise convolution. Moreover, the value branch of this design still maintains the same depth as MLP and GLU, making it backpropagation-friendly. When keeping the parameter volume consistent with the Convolutional Feed-Forward (ConvFFN) [48] with an expansion ratio of R and a convolution kernel size of $k \times k$, the computational complexity of ConvGLU is $2RHWC^2 + \frac{2}{3}RHWCk^2$, which is less than the $2RHWC^2 + RHWCk^2$ of ConvFFN. These attributes render ConvGLU a simple yet more robust mixer, satisfying the diverse requirements of ViTs.

3.3. Architecture Design of TransNeXt

In order to ensure consistency in subsequent ablation experiments 4.2, TransNeXt adopts the same four-stage hierarchical backbone and overlapping patch embedding as PVTv2 [48]. The pooling feature size of the aggregated attention in stages 1-3 is also set to $\frac{H}{32} \times \frac{W}{32}$, identical to PVTv2. In stage 4, as the feature map size has been reduced to $\frac{H}{32} \times \frac{W}{32}$, the feature pooling module cannot function properly. We employ a modified version of multi-head self-ttention (MHSA) that applies query embedding and length-scaled cosine attention. This is consistent with PVTv2's use of MHSA in the fourth stage. For the channel mixer in stages 1-4, we use convolutional GLU with GELU [17] activation. The expansion ratio also follows PVTv2's [8,8,4,4] setting. To ensure consistency with typical MLP parameters, the hidden dimension of convolutional GLU is $\frac{2}{3}\times$ of the set value. Furthermore, we set the head dimension to be 24 for divisibility by 3 in the channel dimension. The specific configurations of TransNeXt variants can be found in appendix.

4. Experiment

Model	#Params. (M)	FLOPs (G)	IN-1K ↑ Top-1(%)	IN-C ↓ mCE(%)	IN-A ↑ Top-1(%)	IN-R ↑ Top-1(%)	Sketch ↑ Top-1(%)	IN-V2 ↑ Top-1(%)
ImageNet-1K 224^2 pre-trained models								
PVT-Tiny [47]	13.2	1.9	75.1	79.6	8.2	33.7	21.3	63.0
PVTv2-B1 [48]	14.0	2.1	78.7	62.6	14.7	41.8	28.9	66.9
BiFormer-T [57]	13.1	2.2	81.4	55.7	25.7	45.4	31.5	70.6
EfficientFormerv2-S2 [29]	12.7	1.3	81.6	–	–	–	–	–
TransNeXt-Micro (Ours)	12.8	2.7	**82.5**	**50.8**	**29.9**	**45.8**	**33.0**	**72.6**
DeiT-Small/16 [42]	22.1	4.6	79.9	54.6	19.8	41.9	29.1	68.4
Swin-T [32]	28.3	4.5	81.2	62.0	21.7	41.3	29.0	69.7
PVTv2-B2 [48]	25.4	4.0	82.0	52.6	27.9	45.1	32.8	71.6
ConvNeXt-T [34]	28.6	4.5	82.1	53.2	24.2	47.2	33.8	71.0
Focal-T [51]	29.1	4.9	82.2	–	–	–	–	–
FocalNet-T (LRF) [52]	28.6	4.5	82.3	55.0	23.5	45.1	31.8	71.2
MaxViT-Tiny [43]	30.9	5.6	83.4	49.6	32.8	48.3	36.3	72.9
BiFormer-S [57]	25.5	4.5	83.8	48.5	39.5	49.6	36.4	73.7
TransNeXt-Tiny (Ours)	28.2	5.7	**84.0**	**46.5**	**39.9**	**49.6**	**37.6**	**73.8**
Swin-S [32]	49.6	8.7	83.1	54.9	32.9	44.9	32.0	72.1
ConvNeXt-S [34]	50.2	8.7	83.1	49.5	31.3	49.6	37.1	72.5
PVTv2-B3 [48]	45.2	6.9	83.2	48.0	33.3	49.2	36.7	73.0
Focal-S [51]	51.1	9.1	83.5	–	–	–	–	–
FocalNet-S (LRF) [52]	50.3	8.7	83.5	51.0	33.8	47.7	35.1	72.7
PVTv2-B4 [48]	62.6	10.1	83.6	46.5	37.1	49.8	37.5	73.5
BiFormer-B [57]	56.8	9.8	84.3	47.2	44.3	49.7	35.3	74.0
MaxViT-Small [43]	68.9	11.7	84.4	46.4	40.0	50.6	38.3	74.0
TransNeXt-Small (Ours)	49.7	10.3	**84.7**	**43.9**	**47.1**	**52.5**	**39.7**	**74.8**
DeiT-Base/16 [42]	86.6	17.6	81.8	48.5	28.1	44.7	32.0	70.9
Swin-B [32]	87.8	15.4	83.5	54.5	35.9	46.6	32.4	72.3
PVTv2-B5 [48]	82.0	11.8	83.8	45.9	36.8	49.8	37.2	73.4
Focal-B [51]	89.8	16.0	83.8	–	–	–	–	–
ConvNeXt-B [34]	88.6	15.4	83.8	46.8	36.7	51.3	38.2	73.7
FocalNet-B (LRF) [52]	88.7	15.4	83.9	49.5	38.3	48.1	35.7	73.5
TransNeXt-Base (Ours)	89.7	18.4	**84.8**	**43.5**	**50.6**	**53.9**	**41.4**	**75.1**
MaxViT-Base [43]	119.5	24.0	**84.9**	43.6	44.2	52.5	40.1	74.5
ImageNet-1K 384^2 fine-tuned models								
Swin-B [32]	87.8	47.1	84.5	–	42.0	47.2	33.4	73.2
ConvNeXt-B [34]	88.6	45.2	85.1	–	45.6	52.9	39.5	75.2
MaxViT-Small [43]	68.9	36.1	85.2	–	48.3	–	–	–
ConvNeXt-L [34]	197.8	101.1	85.5	–	50.7	54.6	41.0	76.0
MaxViT-Base [43]	119.5	74.2	85.7	–	55.1	–	–	–
TransNeXt-Small (Ours)	49.7	32.1	**86.0**	–	**58.3**	**56.4**	**43.2**	**76.8**
TransNeXt-Base (Ours)	89.7	56.3	**86.2**	–	**61.6**	**57.7**	**44.7**	**77.0**

Table 1. A comprehensive comparison on the ImageNet-1K classification and additional robustness test sets.

ImageNet-1K classification: Our code is implemented based on PVTv2 [48] and follows the DeiT [42] recipe for training. The model is trained from scratch on the ImageNet-1K [10] dataset for 300 epochs, leveraging automatic mixed precision (AMP) across $8\times$ GPUs. The specific hyperparameters employed during training are detailed in appendix. To conduct a comprehensive evaluation of the model's robustness, we utilize several additional test sets. These include ImageNet-C [16], a 224^2-sized test set that applies algorithmic distortions to ImageNet-1K validation set; ImageNet-A [18], a test set comprising adversarial examples; ImageNet-R [16], an extended test set containing samples that ResNet-50 [14] failed to classify correctly; ImageNet-Sketch [46], which contains hand-drawn images; and ImageNet-V2 [38], an extended test set that employs the same sampling strategy as ImageNet-1K.

Experimental results: The experimental results, presented in Table 1, establish that our proposed model sets a new benchmark in ImageNet-1K accuracy and robustness across various scales. Specifically, our TransNeXt-Micro model achieves a top-1 accuracy of **82.5%** on ImageNet-1K, surpassing the FocalNet-T(LRF) while utilizing 55% fewer parameters. Similarly, our TransNeXt-Tiny model achieves a top-1 accuracy of **84.0%**, outperforming ConvNeXt-B with a reduction of 69% in parameters. Remarkably, at a resolution of 384^2, our TransNeXt-Small/Base model surpasses the larger MaxViT-Base model by **0.3%/0.5%** respectively after only **5 epochs** of fine-tuning, compared to the 30 epochs used by MaxViT-Base. In terms of robustness, our model exhibits superior performance on five additional test sets. Notably, on the most challenging ImageNet-A test set, TransNeXt demonstrates a significant advantage in robustness as the model scales up. On ImageNet-A at a resolution of 224^2, our TransNeXt-Base surpasses MaxViT-Base by 6.4%. At a resolution of 384^2, our TransNeXt-Small/Base achieves an impressive ImageNet-A accuracy of **58.3%/61.6%**, significantly outperforming ConvNeXt-L by 7.6%/10.9%, while their parameter counts are only 25% and 45% of ConvNeXt-L, respectively.

Object detection and instance segmentation: We employed a Mask R-CNN [15] detection head, trained under a $1\times$ schedule, to evaluate the performance of ImageNet-1K pretrained TransNeXt on object detection and instance segmentation on the COCO [30] dataset. The experimental results are presented in Fig 1. Our model demonstrated comprehensive superiority when compared with previous state-of-the-art models. Notably, even our tiny model surpassed the base models of FocalNet, InternImage and CSWin in terms of AP^b. Similarly, we utilized a DINO [54] detection head, also trained under a $1\times$ schedule, to further assess the potential of our model for object detection. Our TransNeXt-Tiny model achieved an AP^b of 55.1 under a 4-scales setting, surpassing ConvNeXt-L(AP^b of 53.4 in 4-scales setting) 1.7 with only 14% of the latter's backbone parameters. Our TransNeXt-Base achieved an AP^b of 57.1 under a 5-scales setting, approaching the performance of Swin-L(AP^b of 57.2 in 5-scales setting) pretrained on ImageNet-22K.

Semantic segmentation: We used UperNet [50] and Mask2Former [2] methods to train the ImageNet-1K pretrained TransNeXt at a resolution of 512^2 for 160k iterations, and evaluated its semantic segmentation performance on ADE20K [55]. Under the UperNet method, as shown in Fig 1, our TransNeXt demonstrated comprehensive superior-

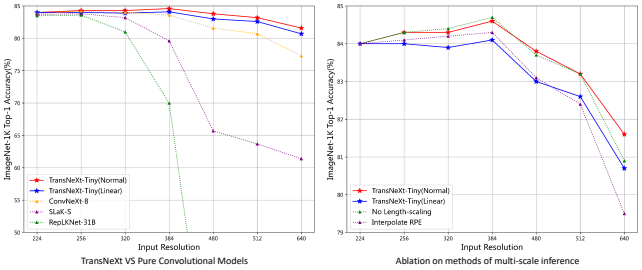

Figure 6. The left figure shows the comparison results of TransNeXt-Tiny under normal and linear inference modes with the pure convolution models on multi-scale image inference performance, while the right figure shows the impact of our positional encoding design and length-scaled cosine attention on this aspect.

ity over previous methods across all sizes. Our TransNeXt-Base even surpassed ConvNeXt-B (mIoU 52.6), which was pretrained on ImageNet-22K and further trained at a resolution of 640^2. Similarly, under the Mask2Former method, our TransNeXt-Small achieved an mIoU of 54.1, surpassing Swin-B (mIoU 53.9) which was pretrained on ImageNet-22K and further trained at a resolution of 640^2. Furthermore, our TransNeXt-Base achieved an mIoU of 54.7. These results indicate that our method has the potential to transcend model size limitations and break through data volume barriers.

Our model demonstrates a even more pronounced performance advantage in dense prediction tasks compared to classification tasks. We believe this validates the effectiveness of the biomimetic vision design of aggregated attention, which enables a more natural visual perception at earlier stages compared to previous methods, as depicted in Fig 2.

4.1. Multi-scale Inference

During inference, TransNeXt in normal mode sets H_p and W_p to $\frac{1}{32}$ of the input size, while in linear mode, these are fixed at 7×7. As depicted in Fig 6 (left), TransNeXt outperforms pure convolutional solutions in both normal and linear modes. Large convolutional kernel schemes [11, 31], also proposed to address depth degradation, exhibit significant performance decline during large image size inference. This reveals the advantage of our approach over large kernel schemes in addressing this issue. For instance, RepLKNet-31B only achieves 0.9% accuracy at a resolution of 640^2. In traditional opinions, pure convolutional models have better multi-scale applicability than ViT models, and such experimental results imply that this opinion needs to be re-examined. The performance decline of large kernel strategies also merits further investigation by the research community.

Fig 6 (right) illustrates the impact of length-scaled cosine and the use of interpolation for position bias on performance. Length-scaling becomes significant at a resolution of 640^2, indicating that sequence length variations exceeding 8× in softmax start to notably diminish the confidence of scaled cosine attention. The application of interpolation for relative position biases results in a substantial performance decline, emphasizing the effectiveness of using extrapolative positional encoding (log-CPB) in multi-scale inference.

4.2. A roadmap from PVT to TransNeXt

Step	Method	#Params. (M)	FLOPs (G)	IN-1K ↑ Top-1(%)	IN-C ↓ mCE(%)	IN-A ↑ Top-1(%)	IN-R ↑ Top-1(%)	Sketch ↑ Top-1(%)	IN-V2 ↑ Top-1(%)
0	PVT-Tiny [47]	13.2	1.9	75.1	79.6	8.2	33.7	21.3	63.0
1	PVTv2-B1 [48]	14.0	2.1	78.7 (+3.6)	62.6 (+17.0)	14.7 (+6.5)	41.8 (+8.1)	28.9 (+7.6)	66.9 (+3.9)
2	Deeper and Thinner	14.9	2.3	80.08 (+1.38)	55.3 (+7.3)	19.7 (+5.0)	43.2 (+1.4)	31.1 (+2.2)	69.2 (+2.3)
3	+ More Heads	14.9	2.3	80.12 (+0.04)	55.0 (+0.3)	19.2 (-0.5)	43.5 (+0.3)	31.5 (+0.4)	69.4 (+0.2)
4	ConvFFN→GLU	14.8	2.2	79.7 (-0.42)	59.5 (-4.5)	18.9 (-0.3)	39.3 (-4.2)	26.8 (-4.7)	69.0 (-0.4)
5	GLU→ConvGLU	14.9	2.2	80.9 (+1.2)	54.6 (+4.9)	23.5 (+4.6)	44.3 (+5.0)	32.7 (+5.9)	70.6 (+1.6)
6	SRA→PFA	12.8	2.7	81.8 (+0.9)	51.7 (+2.9)	26.9 (+3.4)	45.2 (+0.9)	33.3 (+0.6)	71.6 (+1.0)
7	+ Positional Attention	12.8	2.7	82.2 (+0.4)	**50.7** (+1.0)	**31.0** (+4.1)	**46.4** (+1.2)	**34.1** (+0.8)	72.0 (+0.4)
8	+ Query Embedding	12.8	2.7	**82.5** (+0.3)	50.8 (-0.1)	29.9 (-1.1)	45.8 (-0.6)	33.0 (-1.1)	**72.6** (+0.6)

Table 2. The ablation experiments demonstrate the full roadmap from PVT-Tiny to TransNeXt-Micro. In step 1, PVTv2 introduces Overlapping Patch Embedding and Convolutional Feed-Forward (ConvFFN). In step 2, we made PVTv2 consistent with TransNeXt-Tiny in terms of height and width, with a head dimension of 48. In step 3, we decreased the head dimension to 24 and increased the number of attention heads.

Effectiveness of our method: The efficacy of our proposed convolutional GLU (ConvGLU) , pixel-focused attention, positional attention, and query embedding is demonstrated through ablation experiments from step 4 to 8. In the stages of step 4 to 5, step 6, and step 7 to 8, we replaced convolutional feed-forward (ConvFFN) with ConvGLU, spatial-reduction attention (SRA) with pixel-focused attention (PFA), and pixel-focused attention with aggregated attention, respectively. These three substitutions resulted in accuracy improvements of 0.8%, 0.9%, and 0.7% on ImageNet-1K, and 4.3%, 3.4%, and 3.0% on the ImageNet-A test set, respectively, indicating the significant contribution of these three components to performance. It is noteworthy that the introduction of QLV and LKV mechanisms in pixel-focused attention required only an additional 0.2% parameters (from 12.78M to 12.81M) and 0.3% computational overhead (from 2.65G to 2.66G), yet the performance improvement was significant, thereby achieving a cost-effective trade-off. Moreover, in step 4, replacing ConvFFN with GLU led to a significant performance decline, underscoring the necessity of the 3×3 depthwise convolution [4] as conditional position encodings (CPE) [6], particularly as PVTv2's SRA [48] did not use any other positional encoding at this stage. Therefore, step 5 also demonstrated the effectiveness of using ConvGLU as positional encoding.

5. Conclusion

In this work, we propose a biomimetic foveal vision design-based token mixer, **Aggregated Attention**, and a channel mixer with gated channel attention, **Convolutional GLU**. We combine them to propose a powerful and highly robust visual model, **TransNeXt**, which achieves state-of-the-art performance in various visual tasks such as classification, detection, and segmentation. The exceptional performance of TransNeXt in multi-scale inference highlights its advantages over large kernel strategies in addressing the issue of depth degradation. Furthermore, we provide a CUDA implementation that achieves up to 103.4% acceleration in training and 60.5% acceleration in inference. More detailed experimental data and discussions are included in the appendix.

References

[1] Moab Arar, Ariel Shamir, and Amit H. Bermano. Learned queries for efficient local attention. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 10831–10842. IEEE, 2022. 3, 4

[2] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 1280–1289. IEEE, 2022. 7

[3] David Chiang and Peter Cholak. Overcoming a theoretical limitation of self-attention. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 7654–7664. Association for Computational Linguistics, 2022. 5

[4] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 1800–1807. IEEE Computer Society, 2017. 6, 8

[5] Xiangxiang Chu, Zhi Tian, Yuqing Wang, Bo Zhang, Haibing Ren, Xiaolin Wei, Huaxia Xia, and Chunhua Shen. Twins: Revisiting the design of spatial attention in vision transformers. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 9355–9366, 2021. 2

[6] Xiangxiang Chu, Zhi Tian, Bo Zhang, Xinlong Wang, Xiaolin Wei, Huaxia Xia, and Chunhua Shen. Conditional positional encodings for vision transformers. *arXiv preprint arXiv:2102.10882*, 2021. 6, 8

[7] Yann N. Dauphin, Angela Fan, Michael Auli, and David Grangier. Language modeling with gated convolutional networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, pages 933–941. PMLR, 2017. 6

[8] Soham De and Samuel L. Smith. Batch normalization biases residual blocks towards the identity function in deep networks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. 2

[9] Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, Andreas Peter Steiner, Mathilde Caron, Robert Geirhos, Ibrahim Alabdulmohsin, Rodolphe Jenatton, Lucas Beyer, Michael Tschannen, Anurag Arnab, Xiao Wang, Carlos Riquelme Ruiz, Matthias Minderer, Joan Puigcerver, Utku Evci, Manoj Kumar, Sjoerd van Steenkiste, Gamaleldin Fathy Elsayed, Aravindh Mahendran, Fisher Yu, Avital Oliver, Fantine Huot, Jasmijn Bastings, Mark Collier, Alexey A. Gritsenko, Vighnesh Birodkar, Cristina Nader Vasconcelos, Yi Tay, Thomas Mensink, Alexander Kolesnikov, Filip Pavetic, Dustin Tran, Thomas Kipf, Mario Lucic, Xiaohua Zhai, Daniel Keysers, Jeremiah J. Harmsen, and Neil Houlsby. Scaling vision transformers to 22 billion parameters. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, pages 7480–7512. PMLR, 2023. 5

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pages 248–255. IEEE Computer Society, 2009. 7

[11] Xiaohan Ding, Xiangyu Zhang, Jungong Han, and Guiguang Ding. Scaling up your kernels to 31×31: Revisiting large kernel design in cnns. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 11953–11965. IEEE, 2022. 8

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. 1, 3

[13] Michael Hahn. Theoretical limitations of self-attention in neural sequence models. *Trans. Assoc. Comput. Linguistics*, 8:156–171, 2020. 5

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society, 2016. 2, 3, 7

[15] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask R-CNN. *IEEE Trans. Pattern Anal. Mach. Intell.*, 42(2):386–397, 2020. 7

[16] Dan Hendrycks and Thomas G. Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. 7

[17] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016. 7

[18] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 15262–15271. Computer Vision Foundation / IEEE, 2021. 7

[19] Alex Henry, Prudhvi Raj Dachapally, Shubham Shantaram Pawar, and Yuxuan Chen. Query-key normalization for transformers. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, pages 4246–4253. Association for Computational Linguistics, 2020. 5

[20] Jie Hu, Li Shen, Samuel Albanie, Gang Sun, and Enhua Wu. Squeeze-and-excitation networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 42(8):2011–2023, 2020. 6

[21] Md. Amirul Islam, Sen Jia, and Neil D. B. Bruce. How much position information do convolutional neural networks

encode? In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. 6

[22] Bum Jun Kim, Hyeyeon Choi, Hyeonah Jang, Dong Gu Lee, Wonseok Jeong, and Sang Woo Kim. Dead pixel test using effective receptive field. *Pattern Recognit. Lett.*, 167:149–156, 2023. 2

[23] A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. *Master's thesis, Department of Computer Science, University of Toronto*, 2009. 4

[24] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. *Commun. ACM*, 60(6):84–90, 2017. 3

[25] Yann LeCun, Yoshua Bengio, et al. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):1995, 1995. 1, 3

[26] Duo Li, Jie Hu, Changhu Wang, Xiangtai Li, Qi She, Lei Zhu, Tong Zhang, and Qifeng Chen. Involution: Inverting the inherence of convolution for visual recognition. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 12321–12330. Computer Vision Foundation / IEEE, 2021. 3, 4

[27] Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, pages 12888–12900. PMLR, 2022. 4

[28] Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, pages 19730–19742. PMLR, 2023. 4

[29] Yanyu Li, Ju Hu, Yang Wen, Georgios Evangelidis, Kamyar Salahi, Yanzhi Wang, Sergey Tulyakov, and Jian Ren. Rethinking vision transformers for mobilenet size and speed. *CoRR*, abs/2212.08059, 2022. 7

[30] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, pages 740–755. Springer, 2014. 7

[31] Shiwei Liu, Tianlong Chen, Xiaohan Chen, Xuxi Chen, Qiao Xiao, Boqian Wu, Tommi Kärkkäinen, Mykola Pechenizkiy, Decebal Constantin Mocanu, and Zhangyang Wang. More convnets in the 2020s: Scaling up kernels beyond 51x51 using sparsity. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. 8

[32] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 9992–10002. IEEE, 2021. 7

[33] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, and Baining Guo. Swin transformer V2: scaling up capacity and resolution. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 11999–12009. IEEE, 2022. 1, 5

[34] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 11966–11976. IEEE, 2022. 7

[35] Wenjie Luo, Yujia Li, Raquel Urtasun, and Richard S. Zemel. Understanding the effective receptive field in deep convolutional neural networks. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 4898–4906, 2016. 2

[36] Juhong Min, Yucheng Zhao, Chong Luo, and Minsu Cho. Peripheral vision transformer. In *NeurIPS*, 2022. 3

[37] Vinod Nair and Geoffrey E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10), June 21-24, 2010, Haifa, Israel*, pages 807–814. Omnipress, 2010. 5

[38] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, pages 5389–5400. PMLR, 2019. 7

[39] Noam Shazeer. GLU variants improve transformer. *CoRR*, abs/2002.05202, 2020. 6

[40] Jianlin Su. Viewing the scale operation of attention from the perspective of entropy invariance. https://kexue.fm/archives/8823, 2021. 5

[41] Yi Tay, Dara Bahri, Donald Metzler, Da-Cheng Juan, Zhe Zhao, and Che Zheng. Synthesizer: Rethinking self-attention in transformer models. *CoRR*, abs/2005.00743, 2020. 3

[42] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, pages 10347–10357. PMLR, 2021. 7

[43] Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan C. Bovik, and Yinxiao Li. Maxvit: Multi-axis vision transformer. In *Computer Vision - ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXIV*, pages 459–479. Springer, 2022. 1, 2, 7

[44] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 1, 5

[45] Andreas Veit, Michael J. Wilber, and Serge J. Belongie. Residual networks behave like ensembles of relatively shallow networks. In *Advances in Neural Information Processing Systems*

29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain, pages 550–558, 2016. 2

[46] Haohan Wang, Songwei Ge, Zachary C. Lipton, and Eric P. Xing. Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 10506–10518, 2019. 7

[47] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 548–558. IEEE, 2021. 1, 7, 8

[48] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pvtv2: Improved baselines with pyramid vision transformer. *CoRR*, abs/2106.13797, 2021. 4, 6, 7, 8

[49] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. Cvt: Introducing convolutions to vision transformers. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 22–31. IEEE, 2021. 1

[50] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part V*, pages 432–448. Springer, 2018. 7

[51] Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Xiyang Dai, Bin Xiao, Lu Yuan, and Jianfeng Gao. Focal self-attention for local-global interactions in vision transformers. *CoRR*, abs/2107.00641, 2021. 3, 7

[52] Jianwei Yang, Chunyuan Li, Xiyang Dai, and Jianfeng Gao. Focal modulation networks. In *NeurIPS*, 2022. 3, 7

[53] Li Yuan, Qibin Hou, Zihang Jiang, Jiashi Feng, and Shuicheng Yan. VOLO: vision outlooker for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(5):6575–6586, 2023. 3, 4

[54] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M. Ni, and Heung-Yeung Shum. DINO: DETR with improved denoising anchor boxes for end-to-end object detection. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. 7

[55] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ADE20K dataset. *Int. J. Comput. Vis.*, 127(3):302–321, 2019. 7

[56] Daquan Zhou, Zhiding Yu, Enze Xie, Chaowei Xiao, Animashree Anandkumar, Jiashi Feng, and Jose M. Alvarez. Understanding the robustness in vision transformers. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, pages 27378–27394. PMLR, 2022. 6

[57] Lei Zhu, Xinjiang Wang, Zhanghan Ke, Wayne Zhang, and Rynson W. H. Lau. Biformer: Vision transformer with bi-level routing attention. *CoRR*, abs/2303.08810, 2023. 7