Peripheral Vision Transformer

Anonymous Author(s)
Affiliation
Address
email

Abstract

Human vision possesses a special type of visual processing systems called *peripheral vision*. Partitioning the entire visual field into multiple contour regions based on the distance to the center of our gaze, the peripheral vision provides us the ability to perceive various visual features at different regions. In this work, we take a biologically inspired approach and explore to model peripheral vision in deep neural networks for visual recognition. We propose to incorporate peripheral position encoding to the multi-head self-attention layers to let the network learn to partition the visual field into diverse peripheral regions given training data. We evaluate the proposed network, dubbed PerViT, on the large-scale ImageNet dataset and systematically investigate the inner workings of the model for machine perception, showing that the network learns to perceive visual data similarly to the way that human vision does. The state-of-the-art performance in image classification task across various model sizes demonstrates the efficacy of the proposed method.

1 Introduction

For the last ten years, convolution has been a dominant feature transformation in neural networks for visual recognition due to its superiority in modelling spatial configurations of images [17, 25, 26]. Despite the efficacy in learning visual patterns, the local and stationary nature of convolutional kernels limited the maximum extent of representation ability in flexible processing, *e.g.*, dynamic transformations with global receptive fields. Originally devised for natural language processing (NLP), self-attention [46] shed a light on this direction; equipped with adaptive input processing and the ability to capture long-range interactions, it has emerged as an alternative feature transform for computer vision, being widely adopted as a core building block [15]. The stand-alone self-attention models, *e.g.*, ViT [15], however, demand significantly more training data [41] for competitive performance with its convolutional counterparts [5, 19, 52, 21] since they miss certain desirable property which convolution possesses, *e.g.*, locality. These inherent pros and cons of convolution and self-attention encourage recent researches toward combinations of both so as to enjoy the best of the both worlds but which one suits the best for effective visual processing is yet controversial in literature [7, 8, 9, 11, 27, 29, 31, 32, 36, 37, 43, 45, 48, 49, 51, 53, 57].

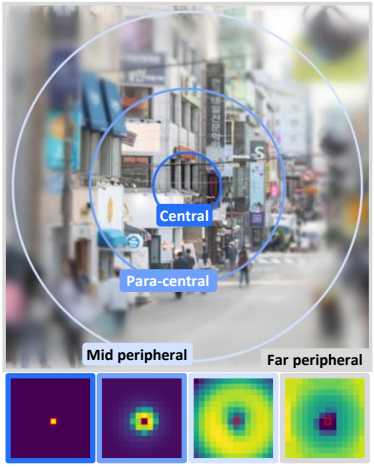

Figure 1: This work explores blending human peripheral vision (top) with attention-based neural networks (bottom) for visual recognition.

Unlike the dominant visual feature transformations in machine vision, human vision possesses a special type of visual processing systems called *peripheral vision* [28]; it partitions the entire visual

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

field into multiple contour regions based on the distances to the center of a gaze where each region identifies different visual aspects. As seen in Fig. 1, we have high-resolution processing near the center of our gaze, *i.e.*, central and para-central regions, to identify highly-detailed visual elements such as geometric shapes, and low-level details. For the regions more distant from the gaze, *i.e.*, mid and far peripheral regions, the resolution decreases to recognize abstract visual features such as motion, and high-level contexts. This systematic strategy enables us to effectively perceive important details within a small fraction (1%) of the visual field while minimizing unnecessary processing of background clutter in the rest (99%), thus facilitating efficient visual processing for human brain.

According to recent study on inner workings of vision transformers [11, 15, 36, 42, 51], their behavior is in fact closely related to how the peripheral vision functions. The attention maps of early layers are learned to locally capture *fine-grained geometric details* at *central regions* while those of later layers perform global attentions to identify *coarse-grained semantics and contexts* from the whole visual field, covering *peripheral regions*. These findings reveal that imitating biological designs may potentially help in modelling an effective machine vision, and also support recent approaches towards a hybrid [4, 9, 18, 11, 30] of convolution and self-attention beyond stand-alone visual processing to take advantages of the two different perception strategies: fine-grained/local and coarse-grained/global, similarly to human vision as illustrated in Fig. 1.

In this work, we take a biologically inspired approach and proposes to inject the peripheral inductive biases to deep neural networks for image recognition. We propose to incorporate peripheral attention mechanism to the multi-head self-attention [46] to let the network learn to partition the visual field into diverse peripheral regions given training data where each region captures different visual features. We experimentally show that the proposed network models effective visual periphery for reliable visual recognition. Our main contributions can be summarized as follows:

- This work explores to narrow the gap between human and machine vision by injecting peripheral inductive biases to self-attention layers, and presents a new form of feature transformations named **M**ulti-head **P**eripheral **A**ttention (MPA).
- Based on the MPA, we introduce **Per**ipheral **Vi**sion **T**ransformer (PerViT), and systematically study the inner workings of PerViT by qualitatively and quantitatively analyzing its learned attentions, which reveal that the network learns to perceive visual elements similarly to the way that human vision does without any special supervisions.
- The state-of-the-art performance on image classification task across different model sizes verify the effectiveness of the proposed approach.

2 Related Work

Feature transformations in computer vision. With notable success in NLP, Transformers [13, 46] introduced a paradigm shift in computer vision [2, 3, 6, 15, 24, 37, 39, 40, 42, 43, 45]. Despite their generalization capability, pure Vision Transformers [15] require extensive amount of training data to capture spatial layout of images due to lack of certain desirable property, *e.g.*, locality. This encouraged many recent ViT work to incorporate local inductive biases via distillation of convolutional biases [42], local self-attention [27, 31, 53], a hybridization [9, 11, 29], and augmenting convolutions [7, 48, 49], all of which convey a unified message: "Despite high generalizibility of self-attentions, a sufficient amount of convolutional processing must be incorporated to capture the spatial configurations of images for reliable visual processing."

Position encoding for Transformer. Witnessing the efficacy of position encoding in capturing input structures in NLP [10, 38, 22], recent vision models [15, 37, 50] have begun employing position encodings for images to model spatial structures of images. In particular, relative position encoding (RPE) plays a vital role for the purpose: The work of Cordonnier *et al.* [8] proves that self-attention has close relationship with convolution when equipped with certain form of RPE. Wu *et al.* [50] explores the existing RPE methods used in vision transformers [10, 22, 38, 47] and draws a conclusion that RPEs imposes convolutional processing on vision transformers. Dai *et al.* [9] observe that depthwise convolution and self-attention can naturally be unified via RPEs. While offering promising directions, the previous RPE work, however, is limited in the sense that the focus of RPE utilization is restricted to only local attention, *e.g.*, convolution. In this work, we exploit RPEs to devise an original visual feature transformation which is more general than convolution while being more structured than self-attention via imitation of human visual system: *peripheral vision*.

3 Our Approach

In this section, we introduce the Peripheral Vision Transformer (PerViT) that learns to model peripheral vision for effective image recognition. We first revisit the mathematical formulation of self-attention and then describe how we improve it with peripheral inductive biases.

Background: Multi-Head Self-Attention. The multi-head self-attention (MHSA) [46] with N_h heads performs an attention-based feature transformation by aggregating N_h self-attention outputs:

$$\text{MHSA}(\mathbf{X}) := \underset{h\in[N_h]}{\text{concat}}\big[\text{Self-Attention}^{(h)}(\mathbf{X})\big]\mathbf{W}_{\text{out}} + \mathbf{b}_{\text{out}}, \tag{1}$$

where $\mathbf{X} \in \mathbb{R}^{HW\times D_{\text{emb}}}$ is a set of input tokens and $\mathbf{W}_{\text{out}} \in \mathbb{R}^{N_h D_h \times D_{\text{emb}}}$ and $\mathbf{b}_{\text{out}} \in \mathbb{R}^{D_{\text{emb}}}$ are the transformation parameters. The N_h outputs of self-attention are designed to extract a diverse set of features from the input representation. Formally, the self-attention at head h is defined as

$$\text{Self-Attention}^{(h)}(\mathbf{X}) := \text{Normalize}\big[\Phi^{(h)}(\mathbf{X})\big]\mathbf{V}^{(h)}, \tag{2}$$

where Normalize$[\cdot]$ denotes a row-wise normalization and $\Phi^{(h)}(\cdot) \in \mathbb{R}^{HW\times HW}$ is a function that provides spatial attentions based on content information to aggregate the values $\mathbf{V}^{(h)} = \mathbf{X}\mathbf{W}^{(h)}_{\text{val}}$:

$$\Phi^{(h)}(\mathbf{X}) := \exp(\tau\mathbf{X}\mathbf{W}^{(h)}_{\text{qry}}(\mathbf{X}\mathbf{W}^{(h)}_{\text{key}})^\top) = \exp(\tau\mathbf{Q}^{(h)}, \mathbf{K}^{(h)\top}), \tag{3}$$

using linear projections of $\mathbf{W}^{(h)}_{\text{qry}}, \mathbf{W}^{(h)}_{\text{key}}, \mathbf{W}^{(h)}_{\text{val}} \in \mathbb{R}^{D_{\text{emb}}\times D_h}$ for queries, keys, and values respectively.

3.1 Peripheral Vision Transformer

Based on the formulation of the MHSA, we define **M**ulti-head **P**eripheral **A**ttention (MPA) as

$$\text{MPA}(\mathbf{X}) := \underset{h\in[N_h]}{\text{concat}}\big[\text{Peripheral-Attention}^{(h)}(\mathbf{X}, \mathbf{R})\big]\mathbf{W}_{\text{out}} + \mathbf{b}_{\text{out}}, \tag{4}$$

where $\mathbf{R} \in \mathbb{R}^{HW\times HW\times D_{\text{r}}}$ is the relative position encoding with D_{r} channel dimension. The self-attention in MHSA is now replaced with Peripheral-Attention, consisting of content- and position-based attention functions ($\Phi^{(h)}_{\text{c}}$, $\Phi^{(h)}_{\text{p}} \in \mathbb{R}^{HW\times HW}$), which is defined as

$$\text{Peripheral-Attention}^{(h)}(\mathbf{X}, \mathbf{R}) := \text{Normalize}\left[\Phi^{(h)}_{\text{c}}(\mathbf{X}) \odot \Phi^{(h)}_{\text{p}}(\mathbf{R})\right]\mathbf{V}^{(h)}, \tag{5}$$

where $\odot$ is Hadamard product which mixes the given pair of attentions to provide a mixed attention $\Phi^{(h)}_{\text{a}}(\mathbf{X}, \mathbf{R}) := \Phi^{(h)}_{\text{c}}(\mathbf{X}) \odot \Phi^{(h)}_{\text{p}}(\mathbf{R}) \in \mathbb{R}^{HW\times HW}$. For the content-based attention Φ_{c}, we use exponentiated (scaled) dot-product between queries and keys as in Eqn. 3: $\Phi^{(h)}_{\text{c}}(\mathbf{X}) := \exp(\tau\mathbf{Q}^{(h)}\mathbf{K}^{(h)\top})$. For the position-based attention Φ_{p}, we design a neural network that aims to imitate human visual system, *e.g.*, peripheral vision.

Modelling peripheral vision: a Roadmap. Human visual field can be grouped into several regions based on the Euclidean distances from the center of gaze, each forming ring-shaped region as seen in Fig. 1, where each region captures different visual aspects; the closer to the gaze, the more complex features we process, and further from the gaze, the simpler visual features we perceive. In the context of 2-dimensional attention map $\Phi_*(\cdot)_{\mathbf{q},:} \in \mathbb{R}^{HW}$, we refer the query position $\mathbf{q} \in \mathbb{R}^2$, *i.e.*, the position where feature of our interest lies at for transformation, as the center of gaze, local regions around the query as central/para-central regions, and the rest as mid/far peripheral regions.

Perhaps the simplest approach to divide the visual field into multiple subregions is to perform a single linear projection on the Eucliden distances, *i.e.*, $\Phi^{(h)}_{\text{p}}(\mathbf{R}) = \sigma\big[\mathbf{R}\mathbf{W}^{(h)}_{\text{p}}\big]$ where $\mathbf{W}^{(h)}_{\text{p}} \in \mathbb{R}^{D_{\text{r}}}$, similarly the previous work of Wu *et al.* [50][1]. For straightforward imitation of peripheral vision, we use Euclidean distance for relative position input $\mathbf{R}$ and weigh the distances in D_{r} different ways for the network to learn the mapping in multiple scales: $\mathbf{R}_{\mathbf{q},\mathbf{k},:} := \text{concat}_{r\in[D_{\text{r}}]}[w_{\text{r}} \cdot \mathbf{R}^{\text{euc}}_{\mathbf{q},\mathbf{k}}] \in \mathbb{R}^{D_{\text{r}}}$ where $\{w_{\text{r}}\}_{r\in[D_{\text{r}}]}$ is a set of learnable parameters shared across layers and heads, and $\mathbf{R}^{\text{euc}}_{\mathbf{q},\mathbf{k}} = \|\mathbf{q} - \mathbf{k}\|_2$

[1] Given $\sigma[\cdot] := \exp(\cdot)$, Peripheral-Attention$^{(h)}$ = Normalize$[\exp(\mathbf{Q}^{(h)}\mathbf{K}^{(h)\top}) \odot \exp(\mathbf{R}\mathbf{W}^{(h)}_{\text{p}})]\,\mathbf{V}^{(h)}$ = softmax$(\mathbf{Q}^{(h)}\mathbf{K}^{(h)\top} + \mathbf{R}\mathbf{W}_{\text{rpe}})\,\mathbf{V}^{(h)}$, which is equivalent to the *bias mode* RPE presented in [50].

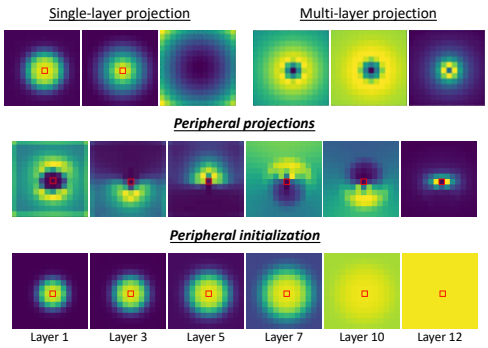

Figure 2: Top: representation ability of Φ_{p} under varying # layers. Middle: peripheral projections. Bottom: peripheral initialization.

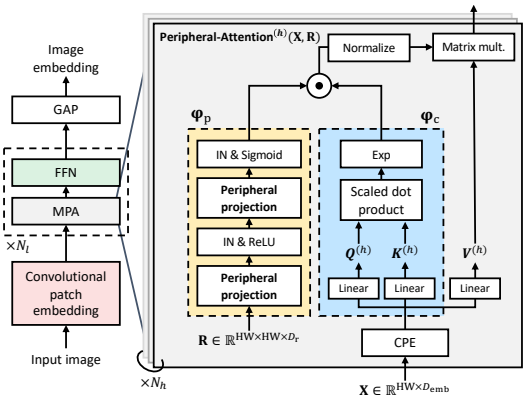

Figure 3: Overall architecture of PerViT which bases on DeiT [11] architecture.

is the Euclidean distance between query and key positions $\mathbf{q}, \mathbf{k} \in \mathbb{R}^2$. For σ, we choose sigmoid to provide normalized (positive) weights to the content-based attention Φ_{c}. A main drawback of this single-layer formulation is that Φ_{p} is only able to provide Gaussian-like attention map as seen in top-left in Fig. 2, thus being unable to represent diverse peripheral regions. For the encoding function to represent diverse (torus-shaped) peripheral regions, the distances must be processed by an MLP:

$$\Phi_{\mathrm{p}}^{(h)}(\mathbf{R}) = \sigma\left[\mathrm{Linear}(\mathrm{ReLU}(\mathrm{Linear}(\mathbf{R}; \mathbf{W}_{\mathrm{p1}}))\mathbf{W}_{\mathrm{p2}}^{(h)})\right] = \sigma\left[\mathrm{ReLU}(\mathbf{R}\mathbf{W}_{\mathrm{p1}})\mathbf{W}_{\mathrm{p2}}^{(h)}\right], \tag{6}$$

where $\mathbf{W}_{\mathrm{p1}} \in \mathbb{R}^{D_{\mathrm{r}} \times D_{\mathrm{hid}}}$ and $\mathbf{W}_{\mathrm{p2}}^{(h)} \in \mathbb{R}^{D_{\mathrm{hid}}}$ are the linear projection parameters[2], and ReLU gives non-linearlity to the function. The first projection $\mathbf{W}_{\mathrm{p1}}$ is shared across the heads in order to exchange information so each function is able to provide attention that are effective or complementary to other heads' attention. Note that given identical relative distances between a fixed query point $\mathbf{q} \in \mathbb{R}^2$ and key points $\mathbf{k}_i, \mathbf{k}_j \in \mathbb{R}^2$, *i.e.*, $\mathbf{R}_{\mathbf{q},\mathbf{k}_i} = \mathbf{R}_{\mathbf{q},\mathbf{k}_j}$, Eq. 6 provides the same attention scores: $\Phi_{\mathrm{p}}(\mathbf{R})_{\mathbf{q},\mathbf{k}_i} = \Phi_{\mathrm{p}}(\mathbf{R})_{\mathbf{q},\mathbf{k}_j}$ as seen in top-right of Fig. 2. This property, however, is not always desired in practical scenarios because the rotational symmetric property hardly holds for most real-world objects. To break the symmetric property in Eq. 6 while preserving peripheral design to sufficient extent, we introduce ***peripheral projections*** in which the transformation parameters are given small spatial resolutions such that $\mathbf{W}_{\mathrm{p1}} \in \mathbb{R}^{K^2 \times D_{\mathrm{r}} \times D_{\mathrm{hid}}}$ and $\mathbf{W}_{\mathrm{p2}}^{(h)} \in \mathbb{R}^{K^2 \times D_{\mathrm{hid}}}$ so the they provide similar but different attention scores, $\Phi_{\mathrm{p}}(\mathbf{R})_{\mathbf{q},\mathbf{k}_i} \neq \Phi_{\mathrm{p}}(\mathbf{R})_{\mathbf{q},\mathbf{k}_j}$, given $\mathbf{R}_{\mathbf{q},\mathbf{k}_i} = \mathbf{R}_{\mathbf{q},\mathbf{k}_j}$, by referring neighboring relative distances around the keys:

$$\Phi_{\mathrm{p}}^{(h)}(\mathbf{R})_{\mathbf{q},\mathbf{k},:} = \sigma\left[\sum_{\mathbf{n}\in\mathcal{N}(\mathbf{k})} \mathrm{ReLU}\left(\sum_{\mathbf{m}\in\mathcal{N}(\mathbf{k})} \mathbf{R}_{\mathbf{q},\mathbf{m},:}\mathbf{W}_{\mathrm{p1}\ \mathbf{m}-\mathbf{k},:,:}\right)\mathbf{W}_{\mathrm{p2}\ \mathbf{n}-\mathbf{k},:}^{(h)}\right], \tag{7}$$

where $\mathcal{N}(\cdot)$ is a function that provides a set of neighbors around input position. After each peripheral projection, we add an instance normalization layer [44] for stable optimization:

$$\mathbf{R}' = \mathrm{ReLU}\left(\mathrm{IN}(\mathrm{PP}(\mathbf{R}; \mathbf{W}_{\mathrm{p1}}); \boldsymbol{\gamma}_{\mathrm{p1}}, \boldsymbol{\beta}_{\mathrm{p1}})\right), \quad \Phi_{\mathrm{p}}^{(h)}(\mathbf{R}) = \sigma\left(\mathrm{IN}(\mathrm{PP}(\mathbf{R}'; \mathbf{W}_{\mathrm{p2}}^{(h)}); \gamma_{\mathrm{p2}}^{(h)}, \beta_{\mathrm{p2}}^{(h)})\right), \tag{8}$$

where $\boldsymbol{\gamma}_{\mathrm{p1}}, \boldsymbol{\beta}_{\mathrm{p1}} \in \mathbb{R}^{D_{\mathrm{hid}}}$ and $\gamma_{\mathrm{p2}}^{(h)}, \beta_{\mathrm{p2}}^{(h)} \in \mathbb{R}$ are weights/biases of the instance norms. The middle row of Fig. 2 depicts learned attentions of Φ_{p} with peripheral projections, which provides peripheral attention maps in greater diversity compared to single- and multi-layer counterparts without $\mathcal{N}$.

Peripheral initialization. Recent study [36, 42] observe that early layers of trained vision transformer learn to attend locally whereas late layers perform global attentions. To facilitate training of our network, we inject this property in the beginning of the training stage by initializing parameters of Φ_{p} for the purpose, making attention scores near the queries larger than the distant ones in the early layers while uniformly distributing them in the late layers as seen in bottom row of Fig. 2. We refer this method as ***peripheral initialization*** for its resemblance to the functioning of peripheral vision [28] which also operates either locally or globally to perceive different visual

[2]We omit the bias terms in the linear layers for brevity.

features [56]. To formally put, given two arbitrarily chosen distances $\mathbf{R}^{\text{euc}}_{\mathbf{q},\mathbf{k}_i}, \mathbf{R}^{\text{euc}}_{\mathbf{q},\mathbf{k}_j} \in \mathbb{R}$ which satisfy $\mathbf{R}^{\text{euc}}_{\mathbf{q},\mathbf{k}_i} < \mathbf{R}^{\text{euc}}_{\mathbf{q},\mathbf{k}_j}$, we want $\Phi_{\text{p}}^{(l,h)}(\mathbf{R})_{\mathbf{q},\mathbf{k}_i} \gg \Phi_{\text{p}}^{(l,h)}(\mathbf{R})_{\mathbf{q},\mathbf{k}_j}$[3] as $l \to 1$, *i.e.*, local attention in early layers, and $\Phi_{\text{p}}^{(l,h)}(\mathbf{R})_{\mathbf{q},\mathbf{k}_i} \approx \Phi_{\text{p}}^{(l,h)}(\mathbf{R})_{\mathbf{q},\mathbf{k}_j}$ as $l \to N_l$, *i.e.*, global attention in late layers, where N_l is the total number of MPA layers. We first initialize the parameters of $\Phi_{\text{p}}^{(l,h)}$ and $\{w_r\}_{r\in[D_r]}$ to particular values. Specifically, for all layers $l \in [N_l]$ and heads $h \in [N_h]$,

$$w_r \coloneqq -c_1, \forall r \in [D_{\text{r}}] \quad \mathbf{W}^{(l)}_{\text{p1}} \coloneqq c_2 J_{K^2,D_{\text{r}},D_{\text{hid}}} \quad \mathbf{W}^{(l,h)}_{\text{p2}} \coloneqq c_2 J_{K^2,D_{\text{hid}}} \quad \boldsymbol{\gamma}^{(l)}_{\text{p1}} \coloneqq \mathbf{1} \quad \boldsymbol{\beta}^{(l)}_{\text{p1}} \coloneqq \mathbf{0} \tag{9}$$

where $c_1, c_2 \in \mathbb{R}^+$ are positive reals, and $J_{N,M} \in \mathbb{R}^{N\times M}$ refers to all-one matrix in size $N \times M$. The above initialization provides local attention after the second peripheral projection, *i.e.*, $\text{PP}(\mathbf{R}'; \mathbf{W}^{(h)}_{\text{p2}})_{\mathbf{q},\mathbf{k}_i,:} > \text{PP}(\mathbf{R}'; \mathbf{W}^{(h)}_{\text{p2}})_{\mathbf{q},\mathbf{k}_j,:}$ given $\mathbf{R}^{\text{euc}}_{\mathbf{q},\mathbf{k}_i} < \mathbf{R}^{\text{euc}}_{\mathbf{q},\mathbf{k}_j}$. Next, based on our findings that biases $\beta^{(l,h)}_{\text{p2}}$ and the weights $\gamma^{(l,h)}_{\text{p2}}$ in the second instance norm control the sizes and strengths of local attention respectively, we simulate peripheral initialization by setting their initial values as $\beta^{(l,h)}_{\text{p2}} \coloneqq s_l$ and $\gamma^{(l,h)}_{\text{p2}} \coloneqq v_l$ where respective $\{s_l\}_{l\in[N_l]}$ and $\{v_l\}_{l\in[N_l]}$ are sets of initial values for attention sizes and strengths. We set their values collected from uniform intervals: $s_l \in [-5.0, 4.0]$ and $v_l \in [3.0, 0.01]$ where $s_{l-1} < s_l$ and $v_{l-1} > v_l$ to give stronger local attentions to shallow layers compared to deep ones as seen in bottom row of Fig. 2. We set $c_1, c_2 = 0.02$ in our experiments and leave all the details on the peripheral initialization to the supplementary.

3.2 Overall Architecture

Based on the proposed peripheral projections and initialization, we develop image classification models, dubbed Peripheral Vision Transformer, which is illustrated in Fig. 3. We follow similar architecture to DeiT [42] with convolutional patch embedding stem; as the original patchify stem [15] exhibits substandard optimizability due to its coarse-grained early visual processing [51], many recent ViT models adopt multi-resolution *pyramidal designs* [31, 48, 49, 53] to mitigate the issue. While the pyramidal models have shown their efficacy in learning reliable image embeddings, we stick with the original single-resolution *cylindrical design* for PerViT because features in multiple resolution make our study less interpretable, which further brings additional technical efforts in combining our peripheral attention Φ_{p} with the existing cost-effective self-attentions mechanisms such as factorized attention [53], shifted-window [31], and cross-shaped window [14]. To carry out fine-grained early processing while keeping single-resolution features across the layers, we adopt convolutional patch embedding layer [51] with multi-stage layouts for channel dimensions. The convolutional embedding layer consists of four 3×3 and one 1×1 convolutions where the 3×3 convolutions are followed by batch norm [23] and ReLU [34]. For additional details, we refer to the supplementary.

Peripheral Vision Transformer. Given an image, the convolutional patch embedding provides token embeddings $\mathbf{X}^{(1)} \in \mathbb{R}^{HW\times D_{\text{emb}}}$. Similarly to [15, 42], the embeddings are fed to a series of N_l blocks each of which consists of an MPA layer and a feed-forward network with residual pathways:

$$\mathbf{X}^{(l')} = \text{MPA}(\text{LN}(\text{CPE}(\mathbf{X}^{(l)}))) + \mathbf{X}^{(l)}, \qquad \mathbf{X}^{(l+1)} = \text{FFN}(\text{LN}(\mathbf{X}^{(l')})) + \mathbf{X}^{(l')}, \tag{10}$$

where LN is layer normalization [1], and FFN is an MLP consisting of two linear transformations with a GELU activation [20]. Following the work of [27, 53], we adopt convolutional position encoding (CPE), *i.e.*, a 3×3 depth-wise convolution, before first layer norm for its efficacy with negligible computational cost. The output $\mathbf{X}^{(N_h)}$ is global-average pooled to form an image embedding.

4 Experiments

In this section, we first investigate the inner workings of PerViT trained on large-scale image classification dataset to examine how it benefits from the proposed peripheral projections and initialization, and then compare the method with previous state of the arts under comparable settings.

Experimental setup. Our experiments focus on image classification on ImageNet-1K [12]. Following training recipes of DeiT [42], we train our model on ImageNet-1K from scratch with batch size of

[3] We now use the superscript to indicate both layer and head indices for the ease of demonstration.

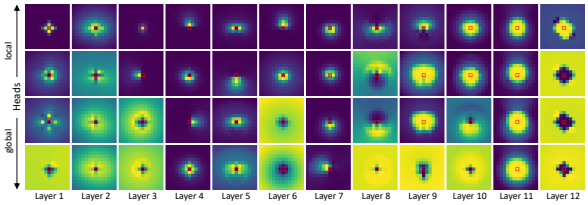

Figure 4: Learned attentions Φ_{p} of PerViT-T.

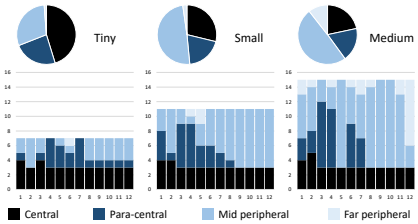

Figure 5: Peripheral region classification.

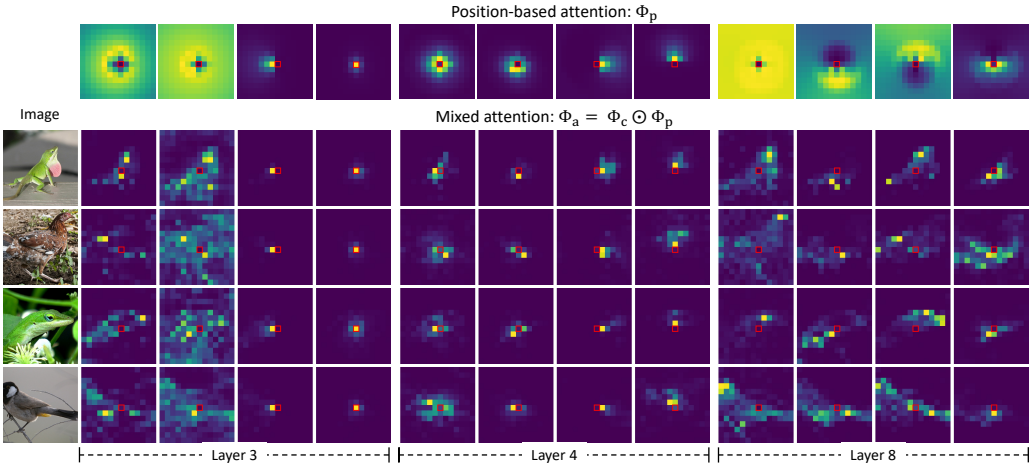

Figure 6: Visualization of learned attentions Φ_{p} and Φ_{a} for $l \in \{3, 4, 8\}$. Best viewed in electronics.

1024, learning rate of 0.001 using AdamW [33] optimizer, cosine learning rate decay scheduler, and the same data augmentations [13] for 300 epochs, including 5 warm-up epochs. We evaluate our model with three different sizes, *e.g.*, Tiny (T), Small (S), and Medium (M). We use stochastic depths of 0.0, 0.1, and 0.2 for T, S, and M respectively. Refer to the supplmentary for additional details.

4.1 The inner workings of PerViT

Learning peripheral vision. We begin by investigating how PerViT models peripheral vision by qualitatively analyzing its learned attention of Φ_{p}. Figure 4 depicts the learned attention map of $\Phi_{\text{p}}^{(l,h)}{}_{\mathbf{q},:} \in \mathbb{R}^{HW}$ for all layers and heads where the query position is given at the center, *i.e.*, $\mathbf{q} = [7, 7]^\top$[4]. We observe that the attentions are learned to be in diverse shapes of peripheral regions. Interestingly, without any special supervisions, the four attended regions ($N_h = 4$) in most layers are learned to complement each other to cover the entire visual field, capturing different visual aspects at each region (head), similarly to human peripheral vision illustrated in Fig. 1. For example, first two heads in Layer 3 attend the central regions while the others cover the rest peripheral regions. The second and third heads in Layer 8 cover top and bottom hemicircles respectively, forming a circular-shaped semi-global receptive field. Moreover, a large number of early attentions is in form of central/para-central regions while those of late layers are learned to cover mid to far peripheral regions. To quantitatively inspect how PerViT models the peripheral visual system, we classify every feature transformation layer in the network into one of the four visual regions $\mathbb{P} \in \{\text{c}, \text{p}, \text{m}, \text{f}\}$ where respective elements refer to central, para-central, mid, and far peripheral regions. PerViT-Attention of head h at layer l is classified as peripheral region p if the average of its attention scores which fall in visual region p is the largest among the others:

$$\text{PeripheralRegion}(l, h) := \arg\max_{p \in \mathbb{P}} \left[\frac{1}{|\mathcal{P}|^2} \sum_{(\mathbf{q},\mathbf{k}) \in \mathcal{P} \times \mathcal{P}} \Phi_{\text{p}}^{(l,h)}{}_{\mathbf{q},\mathbf{k}} \cdot \mathbb{1} \left[\|\mathbf{q} - \mathbf{k}\|_2 \in \mathcal{I}_p \right] \right], \tag{11}$$

[4]The cylindrical design of PerViT provides identical spatial resolution for every intermediate feature map in the network: $H, W = 14$, thus facilitating the ease of qualitative/quantitative analyses of the learned attentions.

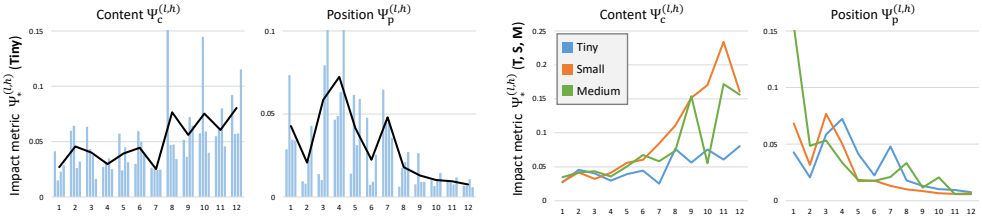

Figure 7: The measure of impact (x-axis: layer index, y-axis: the impact metric Ψ_*). Each bar graph shows the measure of a single head (4 heads at each layer), and the solid lines represent the trendlines which follow the average values of layers. (left: results of PerViT-T. right: results of T, S, and M.)

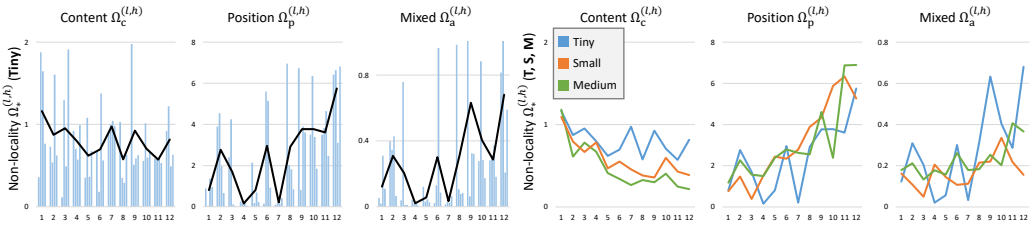

Figure 8: The measure of nonlocality (x-axis: layer index, y-axis: the nonlocality metric Ω_*).

where $\mathcal{P}$ is a set of spatial positions ($|\mathcal{P}| = HW$) and $\mathcal{I}_p$ is distance range (real-valued interval) of peripheral region p[5]. The pie charts of Fig. 5 describe the proportions of peripheral regions for Tiny, Small, and Medium models where the bar graphs show them in layer-wise manner[6]. Similarly to the visualized attention maps in Fig. 4, the early layers attends central/para-central regions whereas deeper ones focus on outer region. We observe that, as the model size grows, the number of mid/far peripheral attention increases whereas that of central/para-central attention stays similar, suggesting that the models no longer require local attentions once sufficient amount of processing is done in the central region because, we hypothesize, identifying geometric patterns, *e.g.*, corners and edges, is relatively simpler process than understanding high-level semantics.

Inspecting the impact of attentions (static *vs.* dynamic). To study how position-based attentions Φ_p contribute to the mixed attentions $\Phi_\mathrm{a} = \Phi_\mathrm{c} \odot \Phi_\mathrm{p}$, we collect sample images and visualize their attention maps of Layers 3, 4 and 8 in Fig. 6. The mixed attentions Φ_a at Layer 4 are formed dynamically (Φ_c) within statically-formed region (Φ_p) while the attentions Φ_a at Layer 8 weakly exploit position information (Φ_p) to form dynamic attentions (Φ_c). The results reveal that Φ_p plays two different roles; it imposes ***semi-dynamic attention*** if the attended region is focused in a small area whereas it serves as ***position bias injection*** when the attended region is relatively broad. In the supplementary, we constructively prove that *an MPA layer in extreme case of semi-dynamic attention/position bias injection is in turn convolution/multi-head self-attention*, naturally generalizing the both transformations. To quantitatively examine the contributions of Φ_c and Φ_p to the mixed attention Φ_a, we define a measure of 'impact' by taking inverse of difference between two attentions:

$$\Psi_\mathrm{p}^{(l,h)} := [||\Phi_\mathrm{a}^{(l,h)} - \Phi_\mathrm{p}^{(l,h)}||_F]^{-1}, \tag{12}$$

where $\|\cdot\|_F$ is Frobenius norm. The higher the measure $\Psi_\mathrm{p}^{(l,h)}$, the larger the impact of position-based attention $\Phi_\mathrm{p}^{(l,h)}$. Being averaged over all test samples, $\Psi_\mathrm{c}^{(l,h)}$ is similarly defined. As seen in Fig. 7, we observe a clear tendency that the impact of position-based attention is significantly higher in early processing, transforming features *semi-dynamically*, while the later layers require less position information, regarding Φ_p as a minor *position bias*. This tendency becomes more visible with larger models as seen in right of Fig 7; Small and Medium models exploit dynamic transformations much more compared to Tiny model, especially in the later layers. Moreover, we note that the impact measures of four heads (bar graphs) within each layer are unevenly distributed, showing high variance, which imply that the network evenly utilizes both position and content information simultaneously within each MPA layer as seen in Layer 3 in Fig. 6, *performing both static/local and dynamic/global*

[5]We use $\mathcal{I}_\mathrm{c} = [0, 1.9)$, $\mathcal{I}_\mathrm{p} = [1.9, 3.37)$, $\mathcal{I}_\mathrm{m} = [3.37, 5.83)$, and $\mathcal{I}_\mathrm{f} = [5.83, 7.9)$. We refer the readers to the supplementary materials for the justification on the these interval choices.

[6]We classify the 3×3 depth-wise convolution in CPE and the two linear projections in FFN as central regions as their receptive fields fall in the interval of $\mathcal{I}_\mathrm{c} = [0, 1.9)$.

Table 1: Study on the effect of each component in PerViT.

Reference	Φ_{p}	Φ_{c}	C-stem	CPE	Top-1	Top-5	FLOPs (G)
(a)	✓	✓	✓	✓	78.8	94.3	1.6
(b)	✗	✓	✓	✓	77.3	94.1	1.6
(c)	✓	✗	✓	✓	76.8	93.5	1.6
(d)	✓	✓	✗	✓	77.8	94.0	1.5
(e)	✓	✓	✓	✗	78.1	94.0	1.6
(f)	✗	✓	✗	✓	76.3	93.2	1.5
(g)	✗	✓	✓	✗	76.7	93.3	1.6
(h)	✓	✓	✗	✗	76.5	93.4	1.5
(i)	✗	✓	✗	✗	72.3	93.4	1.5

Table 2: Comparisons between different relative position encodings with DeiT-Tiny [42] as a baseline.

Method	Top-1	FLOPS (G)
DeiT-T [42]	72.2	1.3
+ CPVT [7]	73.4	2.1
+ iRPE [50]	73.7	1.1
+ PPE (ours)	74.4	1.1

transformation in a single-shot. These results reveal that feature transformations for effective visual recognition should not be restricted to be in neither position-only [19, 32] nor content-only [15, 42] design but they should be in the form of a hybrid [9, 11].

Inspecting the locality (local *vs.* global). We further investigate the inner workings of PerViT by quantifying how locally Φ_* attends. Following the work of [11], we define the measure of nonlocality for $\Phi_{\mathrm{p}}^{(l,h)}$ by summing all pair-wise query-key distances weighted by their attention scores:

$$\Omega_{\mathrm{p}}^{(l,h)} := \frac{1}{|\mathcal{P}|^2} \sum_{(\mathbf{q},\mathbf{k})\in\mathcal{P}\times\mathcal{P}} \Phi_{\mathrm{p}\ \mathbf{q},\mathbf{k}}^{(l,h)} \mathbf{R}_{\mathbf{q},\mathbf{k}}^{\mathrm{euc}}. \tag{13}$$

This metrics Ω_{c} and Ω_{a} are similarly defined, being averaged over all test samples. As seen in Fig. 8, we observe a similar trend of locality between Φ_{p} and Φ_{a}, which reveals the position information play more dominant role over the content information in forming spatial attentions (Φ_{a}) for feature transformation. Interestingly, we also observe that content- and position-based attentions behave conversely; Φ_{c} attends globally in early layers, *i.e.*, large scores are distributed over the whole spatial region, while being relatively local in deeper layers. We hypothesize that the proposed Φ_{p} in early layers is trained to effectively suppress unnecessary scores of Φ_{c} at distant positions, thus exploiting only a few relevant ones within its local region of interest. Meanwhile, Φ_{p} at later layers gives Φ_{c} higher freedom in forming the spatial attention Φ_{a} as described in the plots of Fig. 7, which allows the attention scores of Φ_{c} to be clustered in semantically meaningful parts, *e.g.*, eyes of the animals as seen in attentions of the first head at Layer 8 (Fig. 6), which makes Φ_{c} relatively local.

4.2 Quantitative evaluation on ImageNet-1K

Comparison with state of the arts. Table 3 summarizes the results of our method and recent state of the arts. For fair comparison, the baselines used in our comparison are trained using 224×224 input resolution without distillations, and are grouped into either pyramidal or cylindrical ViT based on the network designs, *i.e.*, multi- or single-resolution feature processing, where the results are partitioned according to model sizes within each group. As shown in the bottom section of Tab 3, the proposed method achieves consistent improvements over the state-of-the-art cylindrical ViT methods [11, 16, 42, 55] while showing competitive results to the pyramidal counterparts. We also implement the proposed position-based attention Φ_{p} on DeiT [42] baseline and compare the results with recent state-of-the-art RPE methods. As seen in Tab. 2, the large improvements over the previous RPE methods [7, 50] further verify the efficacy of the proposed peripheral position encoding (PPE).

Ablation study on Φ_{p}. The top section of Tab. 4 reports results of PerViT-T with different parameter initialization methods for Φ_{p} where peripheral denotes the proposed peripheral initialization, conv refers to convolutional initialization such that $s_l = -5.0$ and $v_l = 3.0$ for all $l \in [N_l]$, and rand refers to random initialization for all parameters in Φ_{p}: w_r, $\mathbf{W}_{\mathrm{p1}}^{(l)}$, $\mathbf{W}_{\mathrm{p2}}^{(l,h)}$, $\boldsymbol{\gamma}_{\mathrm{p1}}^{(l)}$, $\boldsymbol{\beta}_{\mathrm{p1}}^{(l)}$, $\gamma_{\mathrm{p2}}^{(l,h)}$, and $\beta_{\mathrm{p2}}^{(l,h)}$. The results show the efficacy of our peripheral initialization which is also supported by the results in Fig. 4 and 8: Φ_{p} provides early local and late global attentions, suggesting that peripheral initialization effectively reduces burden in learning such form of attentions. The bottom section of Tab. 4 studies network designs for Φ_{p} where $\mathcal{N}$ represents the proposed peripheral projection, *i.e.*, projecting input distance representation by referring neighbors $\mathcal{N}$, ML refers to multi-layer design of Φ_{p}, and Euc & Lrn indicate the type of embedding $\mathbf{R}$: Euc is relative Euclidean distances ($\mathbf{R}_{\mathbf{q},\mathbf{k},:} = \mathrm{concat}_{r\in[D_r]}[w_r \cdot \mathbf{R}]$) where Lrn is relative distances between learnable vectors ($\mathbf{R} \in \mathbb{R}^{HW\times HW\times D_r}$). Without $\mathcal{N}$ and ML, we observe consistent accuracy drops for Euc and Lrn

Table 3: Model performance on ImageNet [12].

Model		Size (M)	FLOPs (G)	Top-1 (%)
Pyramidal Vision Transformers (*multi-resolution*)	PVT-T [48]	13	1.9	75.1
	CoaT-Lite-T [53]	5.7	1.6	77.5
	Swin-T [31]	28	4.5	81.3
	CoaT-Lite-S [53]	20	4.0	81.9
	Focal-T [54]	29	4.9	82.2
	Swin-S [31]	50	8.7	83.0
	CoaT-Lite-M [53]	45	9.8	83.6
	Focal-S [54]	51	9.1	83.5
Cylindrical Vision Transformers (*single-resolution*)	DeiT-T [42]	5.7	1.3	72.2
	XCiT-T12/16 [16]	7.0	1.2	77.1
	PerViT-T (ours)	7.6	1.6	78.8
	DeiT-S [42]	22	4.6	79.8
	T2T-ViT$_t$-14 [55]	22	6.1	81.7
	XCiT-S12/16 [16]	26	4.8	82.0
	PerViT-S (ours)	21	4.4	82.1
	DeiT-B [42]	86	18	81.8
	T2T-ViT$_t$-24 [55]	64	15	82.6
	XCiT-S24/16 [16]	48	9.1	82.6
	PerViT-M (ours)	44	9.0	82.9

Table 4: Ablation study on different initialization methods (top section) and network designs (bottom section) for the position-based attention Φ_p.

Init. method for Φ_p	Top-1	Top-5
Peripheral (ours)	78.8	94.3
Conv	78.6	93.8
Rand	78.5	93.6

Network design for Φ_p	Top-1	Top-5
Euc + $\mathcal{N}$ + ML (ours)	78.8	94.3
Euc + ML	77.9	94.0
Euc	78.0	94.0
Lrn + $\mathcal{N}$ + ML	77.8	94.0
Lrn + ML	77.5	93.8
Lrn	77.6	93.8

by 0.8%p and 0.2%p respectively. A sole multi-layer projection hardly improves accuracy but the model performs the best when $\mathcal{N}$ is jointly used, meaning that both need to complement each other to provide diverse attention shapes as in Fig 4. Furthermore, Euc models consistently surpasses Lrn models, implying the Euclidean distance is more straightforward encoding type than learnable vectors in capturing spatial configurations of images.

Ablation study on main components. In Tab. 1, we analyze the impact of each component in PerViT, which is denoted as (a), where C-stem refers to convolutional patch embedding stem[7]. We observe that the proposed attention Φ_p brings consistent gains to models (b, f, g, i) with relative improvements of 1.4∼4.2%p for (a, d, e, h) respectively. Among three main components (Φ_p, C-stem, CPE), Φ_p has the most significant impact on PerViT (a), losing 1.5%p Top-1 accuracy without Φ_p, *i.e.*, model (b). Surprisingly, PerViT without content-based attention Φ_c, model (c), achieves decent performance, almost equalling to the accuracy of PerViT without position-based attention Φ_p, model (b) (-0.5%p). The results verify that the proposed peripheral attention, which achieves comparable level of efficacy to the content-based attention, learns to generate reliable spatial attentions for visual recognition.

5 Scope and Limitations

Despite the interpretability and effectiveness of PerViT, it still leaves much room for improvements. First, PerViT-Attention (Eq. 5) bases on the original self-attention formulation [46], thus directly inheriting its limitations [15, 42], *e.g.*, quadratic complexity w.r.t. input resolution. The computational efficiency could be further improved by approximating low-rank matrices as in [6, 29, 53]. Second, given the ability to process high-resolution input with feasible complexity, the efficacy of PerViT could be improved by adopting multi-resolution pyramidal design following recent trend of ViT designs [14, 27, 29, 31, 48, 49, 53, 54]. Third, the focus of this paper is model development & exploration for image classification task but we believe the proposed idea is broadly generalizable to other vision applications such as object detection and segmentation. We leave this to future work.

6 Conclusion

This paper explores blending human peripheral vision with machine vision for effective visual recognition, and introduces Peripheral Vision Transformer which learns to provide diverse position-based attentions to model peripheral vision using peripheral projections and initialization. We have systematically investigated the inner workings of the proposed network and observed that the network enjoys the benefits of both convolution and self-attention by learning to decide level of the locality & dynamicity in feature transformation, by the network itself given training data. The consistent improvements over the state of the arts on ImageNet across different model sizes and in-depth ablation study confirm the efficacy of the proposed approach.

[7] We increase feature dimensions of the models (c) (without Φ_c) and (f, h, i) (without C-stem) accordingly to make FLOPs comparable to the others (a-i) to ensure the accuracy drops are not simply due to lower FLOPs.

References

[1] J. Ba, J. Kiros, and G. Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 07 2016. 5

[2] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko. End-to-end object detection with transformers. In *Proc. European Conference on Computer Vision (ECCV)*, 2020. 2

[3] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin. Emerging properties in self-supervised vision transformers. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2021. 2

[4] C.-F. R. Chen, R. Panda, and Q. Fan. RegionViT: Regional-to-Local Attention for Vision Transformers. In *International Conference on Learning Representations (ICLR)*, 2021. 2

[5] F. Chollet. Xception: Deep learning with depthwise separable convolutions. *arXiv preprint arXiv:1610.02357*, 2016. 1

[6] K. Choromanski, V. Likhosherstov, D. Dohan, X. Song, A. Gane, T. Sarlos, P. Hawkins, J. Davis, A. Mohiuddin, L. Kaiser, D. Belanger, L. Colwell, and A. Weller. Rethinking attention with performers. In *International Conference on Learning Representations (ICLR)*, 2021. 2, 9

[7] X. Chu, Z. Tian, B. Zhang, X. Wang, X. Wei, H. Xia, and C. Shen. Conditional positional encodings for vision transformers. *arXiv preprint arXiv:2102.10882*, 2021. 1, 2, 8

[8] J.-B. Cordonnier, A. Loukas, and M. Jaggi. On the relationship between self-attention and convolutional layers. In *International Conference on Learning Representations (ICLR)*, 2020. 1, 2

[9] Z. Dai, H. Liu, Q. V. Le, and M. Tan. Coatnet: Marrying convolution and attention for all data sizes. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 1, 2, 8

[10] Z. Dai, Z. Yang, Y. Yang, J. Carbonell, Q. V. Le, and R. Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context. In *Proc. Association for Computational Linguistics (ACL)*, 2019. 2

[11] S. d'Ascoli, H. Touvron, M. Leavitt, A. Morcos, G. Biroli, and L. Sagun. Convit: Improving vision transformers with soft convolutional inductive biases. In *Proc. International Conference on Machine Learning (ICML)*, 2021. 1, 2, 4, 8

[12] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009. 5, 9

[13] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2019. 2, 6

[14] 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, 2021. 5, 9

[15] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021. 1, 2, 5, 8, 9, 13

[16] A. El-Nouby, 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. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 8, 9, 13

[17] K. Fukushima and S. Miyake. Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition. *Competition and cooperation in neural nets*, 1982. 1

[18] H. He, J. Liu, Z. Pan, J. Cai, J. Zhang, D. Tao, and B. Zhuang. Pruning self-attentions into convolutional layers in single path. *arXiv preprint arXiv:2111.11802*, 2022. 2

[19] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1, 8

[20] D. Hendrycks and K. Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2020. 5

[21] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1

[22] Z. Huang, D. Liang, P. Xu, and B. Xiang. Improve transformer models with better relative position embeddings. In *Annual Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020. 2

[23] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015. 5

[24] A. Jaegle, F. Gimeno, A. Brock, A. Zisserman, O. Vinyals, and J. Carreira. Perceiver: General perception with iterative attention. In *Proc. International Conference on Machine Learning (ICML)*, 2021. 2

[25] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2012. 1

[26] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 1998. 1

[27] Y. Lee, J. Kim, J. Willette, and S. J. Hwang. Mpvit: Multi-path vision transformer for dense prediction. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 2, 5, 9

[28] J. Lettvin. On seeing sidelong. *The Sciences*, 16, 07 1976. 1, 4

[29] 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. *arXiv preprint arXiv:2201.04676*, 2022. 1, 2, 9

[30] J. Liu, H. Li, G. Song, X. Huang, and Y. Liu. Uninet: Unified architecture search with convolution, transformer, and mlp. *arXiv preprint arXiv:2110.04035*, 2021. 2

[31] 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 *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2021. 1, 2, 5, 9, 13

[32] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie. A convnet for the 2020s. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 8

[33] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019. 6

[34] V. Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proc. International Conference on Machine Learning (ICML)*, 2010. 5

[35] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems (NeurIPS)*. 2019. 13

[36] M. Raghu, T. Unterthiner, S. Kornblith, C. Zhang, and A. Dosovitskiy. Do vision transformers see like convolutional neural networks? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 1, 2, 4

[37] P. Ramachandran, N. Parmar, A. Vaswani, I. Bello, A. Levskaya, and J. Shlens. Stand-alone self-attention in vision models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 1, 2

[38] P. Shaw, J. Uszkoreit, and A. Vaswani. Self-attention with relative position representations. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2018. 2

[39] A. Srinivas, T.-Y. Lin, N. Parmar, J. Shlens, P. Abbeel, and A. Vaswani. Bottleneck transformers for visual recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[40] R. Strudel, R. Garcia, I. Laptev, and C. Schmid. Segmenter: Transformer for semantic segmentation. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2021. 2

[41] C. Sun, A. Shrivastava, S. Singh, and A. Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2017. 1

[42] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jegou. Training data-efficient image transformers & distillation through attention. In *Proc. International Conference on Machine Learning (ICML)*, 2021. 2, 4, 5, 8, 9, 13

[43] H. Touvron, M. Cord, A. El-Nouby, P. Bojanowski, A. Joulin, G. Synnaeve, J. Verbeek, and H. Jegou. Augmenting convolutional networks with attention-based aggregation. *arXiv preprint arXiv:2112.13692*, 2021. 1, 2

[44] D. Ulyanov, A. Vedaldi, and V. Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2017. 4

[45] A. Vaswani, P. Ramachandran, A. Srinivas, N. Parmar, B. Hechtman, and J. Shlens. Scaling local self-attention for parameter efficient visual backbones. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1, 2

[46] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 1, 2, 3, 9

[47] H. Wang, Y. Zhu, B. Green, H. Adam, A. Yuille, and L.-C. Chen. Axial-deeplab: Stand-alone axial-attention for panoptic segmentation. In *Proc. European Conference on Computer Vision (ECCV)*, 2020. 2

[48] 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 *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2021. 1, 2, 5, 9, 13

[49] H. Wu, B. Xiao, N. Codella, M. Liu, X. Dai, L. Yuan, and L. Zhang. Cvt: Introducing convolutions to vision transformers. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2021. 1, 2, 5, 9

[50] K. Wu, H. Peng, M. Chen, J. Fu, and H. Chao. Rethinking and improving relative position encoding for vision transformer. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2021. 2, 3, 8

[51] T. Xiao, P. Dollar, M. Singh, E. Mintun, T. Darrell, and R. Girshick. Early convolutions help transformers see better. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 1, 2, 5

[52] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He. Aggregated residual transformations for deep neural networks. *arXiv preprint arXiv:1611.05431*, 2016. 1

[53] W. Xu, Y. Xu, T. Chang, and Z. Tu. Co-scale conv-attentional image transformers. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2021. 1, 2, 5, 9, 13

[54] 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. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 9, 13

[55] L. Yuan, Y. Chen, T. Wang, W. Yu, Y. Shi, Z.-H. Jiang, F. E. Tay, J. Feng, and S. Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2021. 8, 9

[56] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *Proc. European Conference on Computer Vision (ECCV)*, 2014. 5

[57] Y. Zhao, G. Wang, C. Tang, C. Luo, W. Zeng, and Z.-J. Zha. A battle of network structures: An empirical study of cnn, transformer, and mlp. *arXiv preprint arXiv:2108.13002*, 2021. 1

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Section 5.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] We describe potential negative societal impacts in the supplementary.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] We describe the full set of assumptions of theoretical results in the supplementary.
   (b) Did you include complete proofs of all theoretical results? [Yes] We include complete proofs of our theoretical results in the supplementary.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] We include the code and instructions for reproduction in the supplementary. The training and validation data, *e.g.*, ImageNet (ILSVRC2012), is available online.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] We specify all the training details in the supplementary.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No] Training large-scale image classification models is very expensive so we only performed one run. This is common practice in the field as seen in cited references [15, 16, 31, 42, 48, 53, 54].
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] We include the amount and type of resources used for training in the supplementary.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] As our baseline code, we use DeiT [42] which is implemented with PyTorch [35] framework, all of which are open-sourced. We cite all the papers that are relevant to the code, models, and datasets.
   (b) Did you mention the license of the assets? [Yes] Our code refers to the licenses of the assets it relies on.
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] We include trained models in the supplementary for reproduction of our main results.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [No] We are using only publicly available, benchmark datasets.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [No] We are using only publicly available, benchmark datasets.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A] We did not croudsource any information for this research.
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A] We did not croudsource any information for this research.
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A] We did not croudsource any information for this research.