

DAT++: Spatially Dynamic Vision Transformer with Deformable Attention

Zhuofan Xia, Xuran Pan, Shiji Song, Li Erran Li, and Gao Huang

Abstract—Transformers have shown superior performance on various vision tasks. Their large receptive field endows Transformer models with higher representation power than their CNN counterparts. Nevertheless, simply enlarging the receptive field also raises several concerns. On the one hand, using dense attention in ViT leads to excessive memory and computational cost, and features can be influenced by irrelevant parts that are beyond the region of interests. On the other hand, the handcrafted attention adopted in PVT or Swin Transformer is data agnostic and may limit the ability to model long-range relations. To solve this dilemma, we propose a novel deformable multi-head attention module, where the positions of key and value pairs in self-attention are adaptively allocated in a data-dependent way. This flexible scheme enables the proposed deformable attention to dynamically focus on relevant regions while maintains the representation power of global attention. On this basis, we present **Deformable Attention Transformer (DAT)**, a general vision backbone efficient and effective for visual recognition. We further build an enhanced version DAT++. Extensive experiments show that our DAT++ achieves state-of-the-art results on various visual recognition benchmarks, with 85.9% ImageNet accuracy, 54.5 and 47.0 MS-COCO instance segmentation mAP, and 51.5 ADE20K semantic segmentation mIoU.

Index Terms—Vision Transformer, deformable attention, dynamic neural networks.

1 INTRODUCTION

TRANSFORMER [1] is originally introduced to solve natural language processing tasks. It has recently shown great potential in the field of computer vision [2], [3], [4]. The pioneering work, Vision Transformer [2] (ViT), stacks multiple Transformer blocks to process non-overlapping image patch (*i.e.* visual token) sequences, leading to a family of convolution-free models for visual recognition. Compared to their CNN counterparts [5], [6], [7], [8], [9], [10], [11], [12], ViTs have larger receptive fields and excel at modeling long-range dependencies, which are proven to achieve superior performance in the regime of a large amount of training data and model parameters. However, superfluous attention in visual recognition is a double-edged sword and has multiple drawbacks. Specifically, the excessive number of keys to attend per query patch yields high computational cost and slow convergence, and also increases the risk of overfitting.

In order to avoid excessive attention computation, existing works [3], [4], [14], [15], [16], [17], [18], [19], [20], [21], [22] have leveraged carefully designed efficient sparse attention patterns to reduce computational complexity. As two representative approaches among them, Swin Transformer [3] adopts window-based local attention to restrict attention in local windows and shifts windows layer-wise to interact between windows, while Pyramid Vision Transformer (PVT) [4] spatially downsamples the key and value feature maps to save computation by attending queries to a coarsened set of keys. Although effective, hand-crafted attention patterns are data-agnostic and may not be optimal

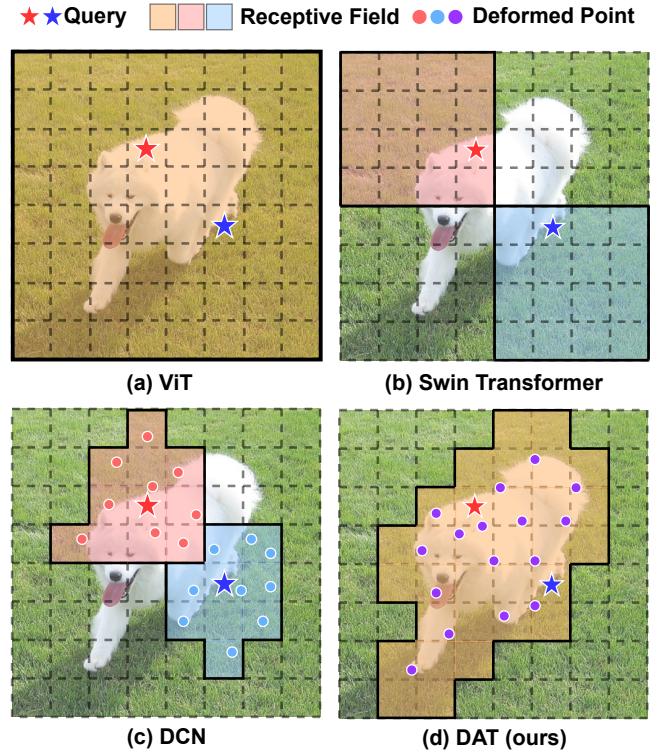


Fig. 1. Comparison of DAT with other Vision Transformers and DCN [13]. The red star and the blue star denote the different queries, and masks with solid line boundaries depict the regions to which the queries attend. In a **data-agnostic** way: (a) ViT [2] adopts full global attention for all queries. (b) Swin Transformer [3] uses partitioned window attention. In a **data-dependent** way: (c) DCN learns different deformed points for each query. (d) DAT learns shared deformed key locations for all queries.

- Zhuofan Xia, Xuran Pan, Shiji Song and Gao Huang are with the Department of Automation, BNRist, Tsinghua University, Beijing 100084, China. Gao Huang is also with Beijing Academy of Artificial Intelligence (BAAI). Corresponding author: Gao Huang.
E-mail: xzf23@mails.tsinghua.edu.cn.
- Li Erran Li is with AWS AI, Amazon, Santa Clara, CA 95054.
- Code is publicly available at <https://github.com/LeapLabTHU/DAT>.

for the various images. It is likely that relevant keys/values are dropped while less important ones are still kept.

Ideally, one would expect that the candidate key/value set for a given query is flexible and has the ability to adapt to each individual input, such that the issues in hand-crafted sparse attention patterns can be alleviated. In fact, in the literature on CNNs, learning deformable receptive fields for the convolution filters has been shown effective in selectively attending to more informative regions on a data-dependent basis, as depicted in Figure 1(c). The most notable work, Deformable Convolution Networks [13], has yielded impressive results on many challenging vision tasks. This motivates us to explore a deformable attention mechanism in Vision Transformers. However, a naive implementation of this idea leads to an unreasonably high memory/computation complexity: the overhead introduced by the deformable offsets is quadratic *w.r.t* the number of visual tokens, prohibiting the applications on high-resolution feature maps by huge memory footprints. As a consequence, although some recent works [23], [24], [25] have investigated the idea of deformable computation in Transformers, none of them has treated it as a fundamental building block to construct a powerful backbone network like the DCN, due to the high computational cost. Among these works, the deformable mechanism is either adopted in the detection head [23], or used as a pre-processing module to sample patches for the subsequent backbone network [24], [25].

In this paper, we present a simple and efficient deformable multi-head attention module (DMHA), equipped with which a powerful pyramid vision backbone, named **Deformable Attention Transformer (DAT)**, is constructed for various visual recognition tasks, including image classification, object detection and instance segmentation, and semantic segmentation, *etc*. Different from DCN, which learns different offsets for different pixels in the whole feature map, we propose to learn a few groups of **query agnostic** offsets, shifting keys and values to capture important regions (as illustrated in Figure 1(d)), based on the observation in [26], [27], [28] that global attention usually results in almost the same attention patterns for different queries. This design introduces a data-dependent sparse attention pattern to Vision Transformer backbones while maintaining linear space complexity. Specifically, for each attention module, the reference points are first generated as uniform grids, which are the same across the input data. Then, the offset generation network takes the query features as input and predicts the corresponding offsets for all reference points. In this way, the candidate keys/values are shifted towards important regions, thus augmenting the original multi-head self-attention with higher flexibility and efficiency to capture more informative features. To achieve stronger performance, we adopted several advanced convolutional enhancements to build DAT++, an improved version of DAT.

Extensive experiments on various visual recognition tasks demonstrate that our DAT++ outperforms or achieves competitive performance compared to other state-of-the-art ViTs and CNNs. Specifically, the largest model DAT-B++ achieves 85.9% Top-1 accuracy on ImageNet [29] image classification, 54.5 bbox mAP and 47.0 mask mAP on MS-COCO [30] instance segmentation, and 51.5 mIoU on ADE20K [31] semantic segmentation, exhibiting the superi-

ority of DAT++.

A preliminary version of this work was published in [32]. In this paper, we extend our conference version in the following aspects:

- We simplify the designs and remove several hyper-parameters in the original DMHA of DAT, leading to a more concise and efficient implementation. We remove the constraint range factor on the learned offsets, extend the DMHA module to all stages of the model, and unify the number of deformed keys across stages.
- We propose an enhanced version of DAT by incorporating overlapped patch embedding and downsampling, injecting depth-wise convolution into MLP blocks, and adopting local perception units, which enhance local features and positional information, and benefit the learning of the proposed deformable attention. These improvements lead to DAT++, a spatial dynamic vision foundation model, which achieves state-of-the-art performance in many challenging vision tasks.
- We provide more analytical experiments and visualization to further study the effectiveness of DAT and DAT++. The ablation experiments are extended with the results of object detection and semantic segmentation. In addition to the study of the core components in DMHA, we sketch a roadmap from DAT to DAT++ as shown in Figure 5, to analyze the effect of each modification in detail. We also show the versatility of the proposed DMHA to various local modeling operators, including different types of local attention and convolution. To distinguish our deformable attention from the one in Deformable DETR, we provide comprehensive comparisons both theoretically and empirically. In addition to quantitative analyses, more exquisite visualization are presented to verify the effectiveness of DAT++.

2 RELATED WORKS

2.1 Vision Transformers

Since the introduction of ViT [2], improvements [3], [4], [14], [15], [18], [19], [33], [34], [35], [36], [37], [38], [39], [40] have focused on learning multiscale features with hierarchical architecture and efficient spatial attention mechanisms. These attention mechanisms include windowed attention [3], [15], [35], [41], pooling attention [19], [20], global or region tokens [14], [17], [37], [42], [43], [44], multiscale feature inductive biases [16], [34], [38], [45], [46], [47], and high-resolution architectures [18], [33] similar to HRNet [48]. In addition to hierarchical architecture, there have also been many attempts to enhance isotropic ViT in many ways, *e.g.*, introducing knowledge distillation [27], [49], [50], [51], [52], improving data augmentations [53], [54], [55], exploring self-supervised learning in Contrastive Learning [56], [57], [58], [59] and Mask Image Modeling [60], [61], [62], [63], [64], [65], and scaling ViTs to larger models [66], [67], [68], [69].

In addition to the development of model architectures or learning targets, it is a growing trend that incorporating dynamic neural networks [70] into ViT can better exploit the spatial structure of input images and significantly improve ViT efficiency. Among these algorithms, dynamic length of visual tokens [71], [72], [73], [74], [75], [76], [77] aims to

determine the vital patch tokens in a data-dependent way and reduce the total length by pruning or merging tokens. Apart from dynamic length, another line of research investigates the dynamic mechanisms in attention modeling, such as dividing image space into trees [78], [79], clustering or predicting super tokens [22], [39], [80], [81], sparse bi-level efficient attention [40], and adapting attention window shapes [82], [83]. In this work, we propose a novel dynamic attention mechanism for visual recognition with deformable attention by learning offsets to deform keys and values, encouraging the deformable attention to focus the important regions in the image.

2.2 Convolution-based Models

Recently, convolution-based approaches have been introduced into Vision Transformer models [84], [85], [86], [87], [88], [89]. ConViT [90] develops a convolutional inductive bias alongside the attention branch. CvT [91] adopts overlapped convolutions in patch embedding and projection layers in attention, and CPE [92] proposes a convolutional position embedding. Based on these early discoveries, more comprehensive works, including CoaT [84], CoAtNet [85], and CMT [86], incorporate more convolution modules and achieve better results. Some analyses [93], [94] also point out the relation of convolutions and local attention, inspiring recent efficient designs of local visual attention, such as QnA-ViT [95], NAT [96], [97], and Slide Transformer [89]. In this work, we apply some convolution-based enhancements on top of DAT to build an improved version, DAT++.

In addition to the convolution-enhanced ViT, recent progress of ViT has inspired several novel CNNs with block structures similar to ViT and larger convolution kernel sizes. ConvNeXt [98] adjusts the conventional CNN structure in both micro and macro designs, building a strong CNN model comparable to Swin Transformer. RepLKNet [99] and SLaK [100] take a step towards larger kernel sizes with structural reparameterization, while HorNet [101] presents another method to expand the receptive field by higher-order convolutions with filters in the global frequency domain [102]. InternImage [103] proposes a strong foundation model based on improved deformable convolution, whose largest variant reaches an unprecedented level on MS-COCO object detection. Extensive experiments show DAT++ can achieve better or comparable performance to the state-of-the-art ViT/CNNs on various tasks.

2.3 Deformable CNN and Attention

Deformable convolution [13], [103], [104] is a powerful dynamic convolution mechanism to attend to flexible spatial locations conditioned on the structure of the input image. Recently, it has been applied to Vision Transformers [23], [24], [25], [105]. Deformable DETR [23] improves the convergence of DETR [106] by attending each query to a small number of deformed keys on the top of a CNN backbone. DPT [24], PS-ViT [25], and SparseFormer [105] build deformable modules to refine visual tokens and pass these tokens to the following isotropic ViT. Specifically, DPT proposes a deformable patch embedding module instead of regular downsampling to refine the patches to the next model stages dynamically. PS-ViT introduces a spatial

sampling module before a ViT backbone to sample visual tokens on the crucial parts of the image. SparseFormer improves PS-ViT by first selecting ROIs in the image and then sampling important points in the ROIs to decode latent tokens, which involves a coarse-to-fine manner to promote efficiency. However, none of these works incorporates deformable attention into vision backbones. On the contrary, our deformable attention takes a powerful yet simple design to learn a set of global keys shared among visual tokens and can be adopted as a general backbone for various vision tasks. Our method can also be viewed as a spatial adaptive mechanism, which has been proven effective in various works [107], [108], [109], [110].

3 DEFORMABLE ATTENTION TRANSFORMER

3.1 Preliminaries

We first revisit the attention mechanism in popular Vision Transformers (ViT) [2], as shown in Figure 1(a). Typically, patch tokens in the shape of $H \times W \times C$ are flattened to a sequence $x \in \mathbb{R}^{N \times C}$, whose length is $N = HW$. Taking the patch sequence as input, a vanilla multi-head self-attention (MHSA) block with M attention heads is formulated as

$$q = xW_q, \quad k = xW_k, \quad v = xW_v, \quad (1)$$

$$z^{(m)} = \sigma(q^{(m)}k^{(m)\top}/\sqrt{d})v^{(m)}, \quad m = 1, \dots, M, \quad (2)$$

$$z = \text{Concat}(z^{(1)}, \dots, z^{(M)})W_o, \quad (3)$$

where $\sigma(\cdot)$ denotes the softmax function, and $d = C/M$ is the dimension of each head. We define $z^{(m)}$ as the aggregated attention values at the m -th head and $q^{(m)}, k^{(m)}, v^{(m)} \in \mathbb{R}^{N \times d}$ to be query, key, and value at the m -th attention head, respectively. The query, key, value, and output projections are parameterized by $W_q, W_k, W_v, W_o \in \mathbb{R}^{C \times C}$ with the bias terms omitted. To build a Transformer block, an MLP block with two linear transformations and a GELU activation is usually adopted to provide nonlinearity.

With normalization layers and identity shortcuts, the l -th Vision Transformer block is formulated as

$$z'_l = \text{MHSA}(\text{LN}(z_{l-1})) + z_{l-1}, \quad (4)$$

$$z_l = \text{MLP}(\text{LN}(z'_l)) + z'_l, \quad (5)$$

where LN is Layer Normalization [111]. Since each token in ViT aggregates features from the entire set of tokens on the feature map, the vanilla MHSA enjoys a global receptive field and excels at modeling long-range spatial dependencies. However, global attention yields a high computational cost of $2(HW)^2C$ in terms of time complexity, restricting the computation to small feature maps with a stride of 16. Vanilla ViT also increases the risk of overfitting, requiring either larger-scale pretraining data [2], [69] or more sophisticated data augmentations [49], [55].

Existing popular hierarchical Vision Transformers, notably PVT [4] and Swin Transformer [3], try to address the challenge of excessive attention. PVT modifies global attention by downsampling the keys and values into smaller spatial sizes, whereas Swin Transformer enforces the attention into local windows, as depicted in Figure 1(b). The downsample of the former results in severe information loss, and the shift-window attention of the latter leads to a

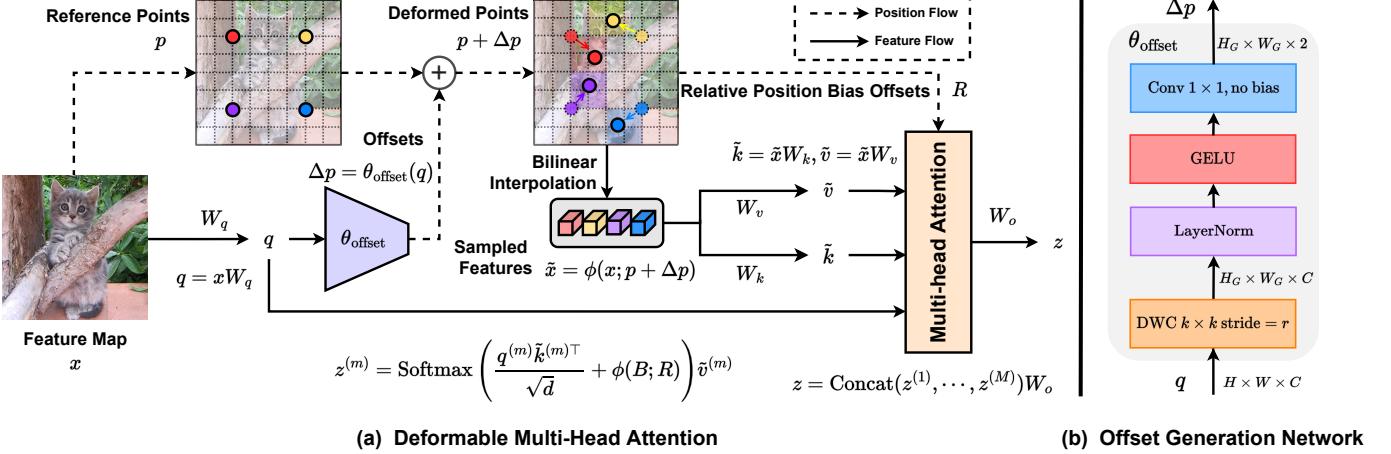


Fig. 2. An illustration of proposed Deformable Attention mechanism. (a) presents the information flow of deformable multi-head attention (DMHA). On the left, a group of reference points is placed uniformly on the feature map, whose offsets are learned from the queries by the offset generation network. The features of important regions are sampled according to the deformed points with bilinear interpolation, which is depicted in the right part. The deformed keys and values are projected by the sampled features and then participate in computing attention. Relative position biases are also computed between the deformed keys and the query grid, enhancing multi-head attention with positional information. We show only 4 reference points for a clear display. (b) reveals the detailed structure of the offset generation network, marked with sizes of feature maps. The query feature map is first transformed by a depth-wise convolution to incorporate local information and then downsampled to the spatial size of the reference points. Another 1×1 convolution following normalization and activation transforms the feature map to the offset values.

much slower growth of receptive fields, which limits the potential of modeling objects lying across multiple windows.

3.2 Deformable Attention

3.2.1 Deformable Convolution and Deformable Attention

The demand for data-dependent sparse attention is hence raised to model spatial relevant features of the tokens more flexibly. A similar methodology is initially proposed by Deformable Convolution Networks [13] in the field of convolutions, which learns the locations of the kernel weights in a data-dependent way. However, migrating the deformable convolution from CNNs to deformable attention Vision Transformers is a nontrivial problem.

In DCN, each element on the feature map learns its offsets individually, as shown in Figure 1(c), of which a 3×3 deformable convolution on an $H \times W \times C$ feature map has a space complexity of $9HWC$. If the same mechanism is applied directly to attention, the space complexity to store queries, keys, and values would drastically rise to $3N_q N_k C$, where N_q, N_k are the numbers of queries and keys which have the same scale as the feature map size HW , bringing approximately a quadratic complexity of $3(HW)^2 C$.

Although Deformable DETR [23] has managed to reduce the overhead by setting a lower number of keys with $N_k = 4$ at each scale and producing attention weights from linear projections of the queries, it is inferior to attend to such few keys with linear attention in a backbone network due to the unacceptable loss of information. The detailed comparison and discussion are included in Sec. 4.5.

In the meantime, the observations in many recent pieces of research [26], [27], [28] reveal that different queries among the visual tokens possess similar attention maps in visual attention models, particularly in the deep layers. Furthermore, only a few keys dominate the attention scores in the above attention maps, indicating that dense attention between queries and keys may be dispensable. Therefore, a simple solution where each query shares the same deformed locations of sparse keys could achieve an efficient trade-off.

3.2.2 Deformable Multi-Head Attention (DMHA)

We herein propose a novel **deformable multi-head attention (DMHA)** mechanism to effectively model the relations among visual tokens under the guidance of learned pivotal regions on the feature maps. These regions are determined by multiple groups of deformed sampling points whose locations are learned from the queries by an offset generation network. For each input image, the deformed points are individually learned to capture the important parts, thus encouraging a data-dependent sparse attention mechanism. By this means, the challenge of excessive attention is tackled along with a relatively low space complexity in that no extra memory is allocated to store keys and values.

The design of DMHA is illustrated in Figure 2(a). From the learned deformed points, the features in the important regions are sampled through bilinear interpolations and then fed to the key and value projections to obtain the deformed keys and values. Finally, the multi-head attention from queries to the sampled keys is applied to aggregate features from the deformed values. Additionally, the locations of deformed points provide a more powerful relative position bias to facilitate the learning of DMHA. These components are discussed in detail below.

Reference points. First, a uniform grid $p \in \mathbb{R}^{H_G \times W_G \times 2}$ is generated as the reference points on the input feature map $x \in \mathbb{R}^{H \times W \times C}$. The grid is downsampled from the input feature map by a factor r , in the size of $H_G = H/r, W_G = W/r$. The reference points are linearly spaced at 2D lattice coordinates in

$$\{(p_x, p_y) | (0, 0), \dots, (W_G - 1, H_G - 1)\}. \quad (6)$$

These locations are then scaled into the range $[-1, +1]$ according to the reference shape $H_G \times W_G$ for numerical stability. At this scale, $(-1, -1)$ indicates the top left corner, and $(+1, +1)$ indicates the bottom right corner.

Offsets learning. To learn the spatial dynamic keys and values conditioned on the queries, we leverage a lightweight

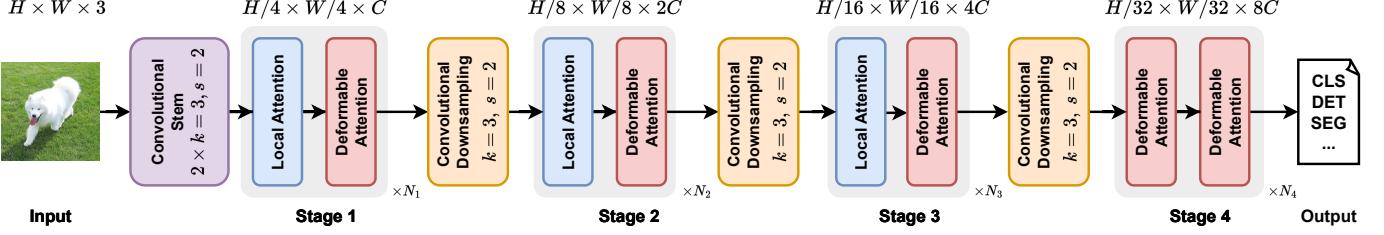


Fig. 3. An illustration of the model architecture of Deformable Attention Transformer++. N_1 to N_4 are the numbers of stacked successive local and deformable attention blocks. k and s denote the kernel size and stride of the convolution layer in patch embeddings. The stem consists of two consecutive layers of convolution-normalization-activation, whereas a normalization layer usually follows the convolution in downsampling blocks.

offset generation network $\theta_{\text{offset}}(\cdot)$ to predict the offsets from query tokens q :

$$\Delta p = \theta_{\text{offset}}(q), \quad (7)$$

whose shape is identical to the reference points in $\Delta p \in \mathbb{R}^{H_G \times W_G \times 2}$. The offsets are added to the positions of the reference points to determine the sampling locations of the deformed points. To avoid sampling outside the feature maps, we clip the sampling locations by -1 and $+1$ to ensure that all the pivotal sampling points lie in the region of the image. Different from the conference version [32], the offset range factor is removed for a concise implementation.

Bilinear sampling. To sample features from the important regions, a sampling function $\phi(\cdot; \cdot)$ with bilinear interpolation is adopted in support of the differentiability:

$$\phi(z; (p_x, p_y)) = \sum_{(r_x, r_y)} g(p_x, r_x)g(p_y, r_y)z[r_y, r_x, :], \quad (8)$$

where $g(a, b) = \max(0, 1 - |a - b|)$ is the bilinear sampling weight function and (r_x, r_y) indexes all the locations on $z \in \mathbb{R}^{H \times W \times C}$. As g would be a nonzero value on the only four lattice points closest to (p_x, p_y) , it simplifies Eq.(8) to a weighted average at the four locations.

Deformable multi-head attention. The features \tilde{x} are sampled with bilinear interpolation at the locations of deformed points. Then the deformed keys and values are obtained by the projections W_k and W_v , respectively. We summarize the above components along with the queries q as follows:

$$q = xW_q, \quad \tilde{k} = \tilde{x}W_k, \quad \tilde{v} = \tilde{x}W_v, \quad (9)$$

$$\text{with } \Delta p = \theta_{\text{offset}}(q), \quad \tilde{x} = \phi(x; p + \Delta p), \quad (10)$$

where \tilde{k} and \tilde{v} represent the deformed key and value embeddings. To achieve the deformable multi-head attention, the queries q are attended to the deformed keys \tilde{k} , modeling the relations between the query features and the pivot regions of the image. The deformed values \tilde{v} are then aggregated under the guidance of the deformable attention map. The mechanism on the m -th attention head is formulated as

$$z^{(m)} = \sigma \left(\frac{q^{(m)} \tilde{k}^{(m)\top}}{\sqrt{d}} + \phi(\hat{B}; R) \right) \tilde{v}^{(m)}, \quad (11)$$

where $\phi(\hat{B}; R) \in \mathbb{R}^{HW \times H_G W_G}$ corresponds to the relative position bias following previous work while with several adaptations to the deformable locations, which will be discussed later in this section. Finally, the aggregated features of each head $z^{(m)}$ are concatenated together and projected through W_o to get the final output z as Eq.(3).

Deformable relative position bias. It is worth noting that the integration of position information into attention enhances the performance of models, including APE [2], RPE [3], CPE [92], LogCPB [41], etc. The relative position embedding (RPE) proposed in Swin Transformer encodes the relative positions between every pair of query and key, which augments the vanilla attention with spatial inductive bias. This explicit modeling on relative locations is best suitable for deformable attention, where the deformed keys have arbitrary continuous locations rather than fixed discrete grids. Following RPE, considering a feature map of shape $H \times W$, its relative coordinate displacements lie in the range of $[-H, +H]$ and $[-W, +W]$ at two spatial dimensions. Thus, an RPB table $\hat{B} \in \mathbb{R}^{(2H-1) \times (2W-1)}$ holding parameters is constructed to determine the deformable relative position bias by looking up the table. Since our deformable attention has continuous key positions, we compute the relative locations in the normalized range $[-1, +1]$ and then sample $\phi(\hat{B}; R)$ with bilinear interpolation in the parameterized bias table \hat{B} by the continuous relative displacements to cover all possible offset values.

Offset generation. As stated above, a lightweight sub-network is adopted for offset generation, which consumes the query features and outputs the corresponding offset values for each reference point. Considering that each reference point covers a local $r \times r$ region, the offset generation network should also have at least $r \times r$ perception of the local features to learn reasonable offsets. Therefore, we design the sub-network as two convolution modules with a combination of normalization and activation, as depicted in Figure 2(b). The input features are first passed through a $k \times k$ depth-wise convolution with r stride to capture local features, where k is slightly larger than r to ensure the reasonable receptive field, followed by a LayerNorm and a GELU activation. Then, a 1×1 convolution is adopted to predict the 2-D offsets, whose bias is dropped to alleviate the inappropriate compulsive shift for all locations.

Offset groups. To promote the diversity of the deformed points, we follow a similar paradigm in MHSA that the token channels are split into multiple heads to compute a variety of attention. Therefore, we split the channels into G groups to generate diverse offsets. The offset generation network shares weights for features from different groups to generate the corresponding offsets. In practice, the head number M of attention is set to be multiple times the size of the offset groups G , guaranteeing that multiple attention heads are assigned to one group of sampled features.

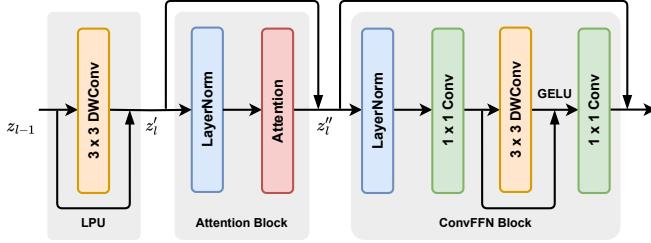


Fig. 4. Detailed architecture of each Transformer block in DAT++, enhanced with Local Perception Unit (LPU) and ConvFFN blocks.

3.2.3 Complexity Analysis

Deformable multi-head attention (DMHA) has a similar computation cost as the counterpart in PVT or Swin Transformer. The only additional overheads come from the offset generation network and the bilinear feature sampling. The time complexity of DMHA can be summarized as:

$$\Omega(\text{DMHA}) = \underbrace{2HWN_s C + 2HWC^2}_{\text{multi-head self-attention}} + \underbrace{2N_s C^2 + (k^2 + 6)N_s C}_{\text{offsets \& sampling}}, \quad (12)$$

where $N_s = H_G W_G = HW/r^2$ is the number of sampled points. It can be immediately seen that the computational cost of the offset network has linear complexity *w.r.t.* the spatial and channel size, which is comparably minor to the cost for attention computation. Besides, the smaller N_s further reduces the computation cost in the projection layers of keys and values by first sampling the features.

We adopt Floating Point Operations (FLOPs) to measure the computational complexity. *E.g.*, the 3rd stage of a hierarchical model with 224×224 input for image classification usually has the computation sizes of $H = W = 14$, $N_s = 49$, $C = 384$, thus the computational complexity for multi-head self-attention in a single DMHA block is 79.63M in FLOPs. Combining the offset generation network, whose additional overhead is 0.58M in FLOPs, which is only 0.7% in 80.21 MFLOPs of the whole module. Additionally, the complexity could be further reduced by enlarging downsample factor r , making it friendly to tasks with much higher resolution inputs such as object detection and instance segmentation.

3.3 Model Architectures

3.3.1 Building Deformable Attention Transformer

We replace the vanilla MHSA (Eq.(2)) with the proposed DMHA (Eq.(11)) in the attention layer (Eq.(4)) and combine it with an MLP (Eq.(5)) to build a Vision Transformer block with deformable attention. Based on these blocks, our model, **Deformable Attention Transformer (DAT)**, shares a similar pyramid structure with [3], [4], [24], [112], which is broadly applicable to various vision tasks requiring multi-scale feature maps. The architecture of DAT is illustrated in Figure 3, in which an input image with shape $H \times W \times 3$ is firstly embedded by a convolutional stem with stride 4 to encode the $H/4 \times W/4 \times C$ patch embeddings. Aiming to build a hierarchical feature pyramid, the backbone includes 4 model stages with a progressively increasing stride at each stage. Between two consecutive stages, there is a convolutional downsampling layer with stride 2. In each stage, sequential building blocks of different types of

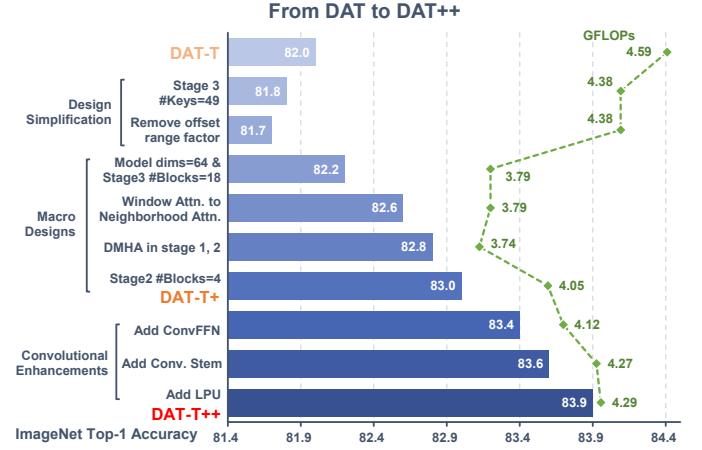


Fig. 5. Roadmap from DAT [32] to the improved DAT++. From the top to the bottom, each row represents a version of model, where the modifications are appended to the model in the last row. The blue bar chart reports the top-1 accuracy of each model on ImageNet next to the green line chart which shows the complexity of each model in GFLOPs.

attention extract features from the patch embeddings by spatial mixing in the attention layers and channel mixing in the MLP layers.

After stages of processing, DAT learns the representations of the input image hierarchically, thus ready for various vision tasks. For the classification task, we first normalize the output from the last stage by LN [111] and then adopt a linear classifier with pooled features to predict the class label. For the object detection, instance segmentation, and semantic segmentation tasks, DAT plays the role of a feature extractor in some well-designed detectors or segmentors. To build the feature pyramid for these tasks, we add an LN layer to the features from each stage before feeding them into the following modules such as FPN [113] for object detection or decoders in semantic segmentation.

3.3.2 Enhancing DAT to DAT++

To fully explore the capacity of DAT, we consider incorporating several enhancements of learning better local features into the architecture to boost performance and build an improved version named DAT++, which is sketched as a roadmap in Figure 5. We introduce these enhancements as follows, and detailed discussions are provided in Sec. 4.4.

Overlapped patch embedding layers provide remarkable gains over their non-overlapped counterparts [2], [3], [4], [49], [50], which are adopted in DAT++ to take the place of the stem layers and downsampling layers, as done in [39], [40], [86], [103]. This overlapped convolution manner could mitigate the severe information loss in the unfolding operation and produce high-quality feature maps, which is essential to learn accurate offsets for DMHA to dynamically focus on the important regions in the image. Specifically, we use the 3×3 convolution with stride 2 to embed image pixels into tokens, and downsample feature maps across stages.

Local Perception Unit (LPU) [86] is a depth-wise convolution wrapped by a residual connection. Similar to CPE [92], this layer is usually placed at the top of every transformer block to enhance positional information implicitly. Since deformable attention is highly sensitive to positional features and the geometric distributions in the image, the implicit

TABLE 1

DAT++ model architecture specifications. N_i : Number of blocks at stage i . C : Base channels in each block. r : Downsample ratio of deformed points. M : Number of attention heads in DMHA. G : Number of offset groups in DMHA. K : Kernel size of local neighborhood attention. The spatial size and stride of feature maps are labeled under each stage.

Variant Architectures of Deformable Attention Transformer			
(224 × 224)	DAT-T++	DAT-S++	DAT-B++
Stage 1 (56 × 56)	$N_1=1, C=64$ $r=8, M=2$ $G=1, K=7$	$N_1=1, C=96$ $r=8, M=3$ $G=1, K=7$	$N_1=1, C=128$ $r=8, M=4$ $G=2, K=7$
Stride: 4			
Stage 2 (28 × 28)	$N_2=2, C=128$ $r=4, M=4$ $G=2, K=7$	$N_2=2, C=192$ $r=4, M=6$ $G=2, K=7$	$N_2=2, C=256$ $r=4, M=8$ $G=4, K=7$
Stride: 8			
Stage 3 (14 × 14)	$N_3=9, C=256$ $r=2, M=8$ $G=4, K=7$	$N_3=9, C=384$ $r=2, M=12$ $G=3, K=7$	$N_3=9, C=512$ $r=8, M=16$ $G=8, K=7$
Stride: 16			
Stage 4 (7 × 7)	$N_4=1, C=512$ $r=1, M=16$ $G=8$	$N_4=1, C=768$ $r=1, M=24$ $G=6$	$N_4=1, C=1024$ $r=1, M=32$ $G=16$
Stride: 32			

conditional position encoding added to feature maps could provide hints of the important regions, which benefits the learning of the offset generation module of DMHA.

ConvFFN adds a depth-wise convolution between two fully connected layers in the FFN (MLP) for channel mixing, reinforcing the spatial modeling ability [21], [86], [99], [114]. ConvFFN serves as a local feature modeling operation, which complements the global attention DMHA with better locality modeling. DAT++ chooses a variant of ConvFFN with a skip connection rather than plain convolution in PVTv2 [21] or additional BatchNorms [115] in CMT [86].

The building block of DAT++ equipped with the above enhancements is illustrated in Figure 3 - 4. In DAT++, we modify the l -th Transformer encoder block (Eq.(4) - (5)) with convolutions injected:

$$z'_l = \text{LPU}(z_{l-1}), \quad (13)$$

$$z''_l = \text{MHSA}(\text{LN}(z'_l)) + z'_l, \quad (14)$$

$$z_l = \text{ConvFFN}(\text{LN}(z''_l)) + z''_l. \quad (15)$$

To make fair comparisons, we carefully ablate these additional components in Sec. 4.4 to verify the effectiveness of our proposed deformable attention.

3.3.3 Architecture Design

We introduce successive local attention and deformable attention blocks in the first three stages of DAT++. Unlike the conference version [32], which leverages window-based attention in Swin Transformer [3] as the local attention layer, we opt for Neighborhood Attention [96] as its efficient local self-attention possesses more representation power.

In the first three stages of DAT++, the feature maps are processed by a transformer block with local attention to aggregate information within a neighbor window for visual tokens and then pass through the deformable attention block to model the global relations between the locally augmented tokens. This alternate design of attention blocks with local and global receptive fields helps the model learn strong representations, sharing similar thoughts in [35], [97],

TABLE 2

Detailed hyper-parameters for training variants of DAT++ on ImageNet. PT and FT means pretraining and finetuning, respectively.

Settings	DAT-T++	DAT-S++	DAT-B++
Input resolution	PT 224 ²	PT 224 ²	PT 224 ² FT 384 ²
Batch size	4096	4096	4096 512
Optimizer	AdamW	AdamW	AdamW AdamW
Learning rate	4×10 ⁻³	4×10 ⁻³	4×10 ⁻³ 2×10 ⁻⁵
LR schedule	cosine	cosine	cosine cosine
Weight decay	5×10 ⁻²	5×10 ⁻²	5×10 ⁻² 1×10 ⁻⁸
Warmup epochs	5	5	5 0
Epochs	300	300	300 30
Horizontal flip	✓	✓	✓ ✓
Random resize & Crop	✓	✓	✓ ✓
AutoAugment	✓	✓	✓ ✓
Mixup alpha	0.8	0.8	0.8 0.8
Cutmix alpha	1.0	1.0	1.0 1.0
Random erasing prob.	0.25	0.25	0.25 0.25
Color jitter	0.4	0.4	0.4 0.4
Label smoothing	0.1	0.1	0.1 0.1
Dropout	✗	✗	✗ ✗
Droppath rate	0.2	0.4	0.6 0.6
Repeated augment	✗	✗	✗ ✗
Grad. clipping	5.0	5.0	5.0 5.0

[116], [117], [118], [119], [120], which has been shown to be an effective design. Since the spatial resolution is rather limited in the last stage, the neighbor size of the local attention is close to the size of feature maps. The local attention performs similarly to the vanilla MHSA. Therefore, all the transformer blocks in the fourth stage of DAT++ are set to deformable attention. We discuss this design in Sec. 4.4.

We build three variants of DAT++ at different scales of parameters and FLOPs for fair comparisons with other ViTs and CNNs. We change the model scale by increasing the hidden dimensions, leading to DAT-T/S/B++. The detailed architectures are summarized in Table 1. Note that there are other design choices for the local modeling layers, e.g., large kernel depth-wise convolutions [98], and window-based local attention [3]. We show that DMHA is versatile and works well with other local operators in Sec. 4.4.

4 EXPERIMENTS

We analyze DAT++ on several representative visual recognition tasks, including ImageNet [29] image classification, MS-COCO [30] object detection, and ADE20K [31] semantic segmentation. In addition to comparing DAT with other leading models on these standard benchmark datasets, we provide detailed ablation experiments and visualization to analyze the effectiveness of the proposed DAT++.

4.1 Image Classification

Settings. ImageNet [29] is a widely used image classification dataset for deep neural networks, which contains 1.28M examples for training and 50K examples for validation. We train three variants of DAT++, including DAT-T/S/B++, on the training split and report the Top-1 accuracy on the validation split to compare with other Vision Transformers and Convolution Neural Networks.

TABLE 6

Object detection and instance segmentation performance on MS-COCO dataset with Cascade Mask R-CNN [131] detector. This table shows the number of parameters, computational cost (FLOPs), mAP at different IoU thresholds, and different object sizes. The FLOPs are computed over the entire detector at the resolution of $1280 \times 800 \times 3$.

Cascade Mask R-CNN Object Detection & Instance Segmentation on MS-COCO															
Method	FLOPs	#Param	Schedule	AP ^b	AP ^b ₅₀	AP ^b ₇₅	AP ^b _s	AP ^b _m	AP ^b _l	AP ^m	AP ^m ₅₀	AP ^m ₇₅	AP ^m _s	AP ^m _m	AP ^m _l
Swin-T [3]	745G	86M	1x	48.1	67.1	52.2	30.4	51.5	63.1	41.7	64.4	45.0	24.0	45.2	56.9
DAT-T [32]	750G	86M	1x	49.1	68.2	52.9	31.2	52.4	65.1	42.5	65.4	45.8	25.2	45.9	58.6
DAT-T++	771G	81M	1x	52.2	70.9	56.6	33.9	56.2	68.1	45.0	68.1	48.9	25.1	48.5	63.4
Swin-T [3]	745G	86M	3x	50.4	69.2	54.7	33.8	54.1	65.2	43.7	66.6	47.3	27.3	47.5	59.0
ConvNeXt-T [98]	741G	86M	3x	50.4	69.1	54.8	33.9	54.4	65.1	43.7	66.5	47.3	24.2	47.1	62.0
NAT-T [96]	737G	85M	3x	51.4	70.0	55.9	33.8	55.2	66.2	44.5	67.6	47.9	23.7	47.8	63.4
DiNAT-T [97]	737G	85M	3x	52.2	71.0	56.9	36.1	55.9	67.9	45.2	68.3	48.9	26.4	48.4	64.5
HorNet-T [101]	728G	80M	3x	52.4	71.6	56.8	36.4	56.8	67.1	45.6	69.1	49.6	26.5	49.2	64.3
CSWin-T [15]	757G	80M	3x	52.5	71.5	57.1	-	-	-	45.3	68.8	48.9	-	-	-
DAT-T [32]	750G	86M	3x	51.3	70.1	55.8	34.1	54.6	66.9	44.5	67.5	48.1	27.9	47.9	60.3
DAT-T++	771G	81M	3x	53.0	71.6	57.7	37.1	56.6	68.6	46.0	69.3	50.1	26.7	49.3	64.3
Swin-S [3]	838G	107M	3x	51.9	70.7	56.3	35.2	55.7	67.7	45.0	68.2	48.8	28.8	48.7	60.6
ConvNeXt-S [98]	827G	108M	3x	51.9	70.8	56.5	35.4	55.7	67.3	45.0	68.4	49.1	25.2	48.4	64.0
NAT-S [96]	809G	108M	3x	51.9	70.4	56.2	35.7	55.9	66.8	44.9	68.2	48.6	25.8	48.5	63.2
DiNAT-S [97]	809G	108M	3x	52.9	71.8	57.5	37.6	56.8	68.3	45.8	69.3	49.8	26.5	49.4	64.7
HorNet-S [101]	827G	108M	3x	53.3	72.3	57.8	36.7	57.8	69.0	46.3	69.9	50.4	26.8	50.2	65.1
CSWin-S [15]	820G	92M	3x	53.7	72.2	58.4	-	-	-	46.4	69.6	50.6	-	-	-
DAT-S [32]	857G	107M	3x	52.7	71.7	57.2	37.3	56.3	68.0	45.5	69.1	49.3	30.2	49.2	60.9
DAT-S++	895G	110M	3x	54.2	72.7	58.9	38.0	58.3	69.7	46.9	70.1	51.3	28.3	50.3	65.8
Swin-B [3]	982G	145M	3x	51.9	70.5	56.4	35.4	55.2	67.4	45.0	68.1	48.9	28.9	48.3	60.4
NAT-B [96]	931G	147M	3x	52.2	70.9	56.8	35.2	55.9	67.2	45.1	68.3	49.1	25.5	48.4	64.0
ConvNeXt-B [98]	964G	146M	3x	52.7	71.4	57.2	36.5	56.6	68.2	45.6	68.9	49.6	26.0	49.0	64.6
DiNAT-B [97]	931G	147M	3x	53.4	72.0	58.2	37.4	57.5	68.8	46.2	69.7	50.2	26.4	49.7	65.1
CSWin-B [15]	1004G	135M	3x	53.9	72.6	58.5	-	-	-	46.4	70.0	50.4	-	-	-
HorNet-B [101]	965G	146M	3x	54.0	72.8	58.7	37.8	58.1	69.7	46.9	70.2	51.1	27.4	50.4	65.7
DAT-B [32]	1003G	145M	3x	53.0	71.9	57.6	36.0	56.8	69.1	45.8	69.3	49.5	29.2	49.5	61.9
DAT-B++	1059G	151M	3x	54.5	73.0	59.4	38.5	58.4	69.8	47.0	70.5	51.4	27.9	50.3	65.8

4.2 Object Detection and Instance Segmentation

Settings. MS-COCO [30] dataset has about 118K training images and 5K validation images, which is a widely accepted benchmark for object detection and instance segmentation. To validate our proposed models on downstream vision tasks, we use DAT++ as a backbone network and incorporate it into the detectors to extract visual features from the images containing various objects and instances. DAT++ is employed in three widely used detectors: RetinaNet [121], Mask R-CNN [130], and Cascade Mask R-CNN [131], and initialized from the ImageNet pretrained model weights with 300 epochs. Following common practice [3], [4], [98], we train object detection and instance segmentation models with DAT++ for 12 (1x schedule) or 36 (3x schedule) epochs.

In practice, we train the models on 32 GPUs with a linearly scaled learning rate of 4×10^{-3} and batch size of 64. Since different training durations usually cause different convergence, we adjust the stochastic depth factor for each model on each scale and schedule to prevent overfitting. For RetinaNet [121], this factor is set to 0.0 (1x), 0.2 (3x) for DAT-T++, and 0.1 (1x), 0.5 (3x) for DAT-S++. For Mask R-CNN [130], the drop path rates are 0.0 (1x), 0.3 (3x) for DAT-T++, and 0.1 (1x), 0.5 (3x) for DAT-S++. For Cascade Mask-RCNN [131], the rates are 0.0 (1x), 0.1 (3x) for DAT-

T++, 0.5 (3x) for DAT-S++, and 0.8 (3x) for DAT-B++.

Results. Table 4 shows the object detection results of DAT-T/S++ on RetinaNet [121] in the 1x and 3x schedules. The *tiny* variant achieves 46.8 mAP on 1x schedule and 49.2 mAP on 3x schedule, surpassing other methods by at least 1.5 mAP, and the *small* variant with respective 48.3 mAP (1x) and 50.2 mAP (3x) exhibits significant improvements over other baselines with similar parameters and FLOPs.

As for instance segmentation, we combine the *Tiny*, *Small* scale variants of DAT++ on Mask R-CNN [130] in the 1x and 3x schedules. As shown in Table 5, DAT-T++ achieves consistent improvements on several SOTA models [15], [103] by up to 1.5 bbox mAP and 1.4 mask mAP. DAT-S++ even outperforms some *Base* variants of the baselines with at least 0.4 bbox mAP and 0.5 mask mAP. Similar performance gains are also observed in Cascade Mask R-CNN [131] framework. Table 6 reports that DAT++ reaches a new level of precision with 53.0 bbox mAP, 46.0 mask mAP of *Tiny*, 54.2, 46.9 of *Small*, 54.6, 47.2 of *Base*, respectively, which show promising margins over other strong ViT/CNN models.

4.3 Semantic Segmentation

Settings. ADE20K [31] is a popular dataset for semantic segmentation with 20K training examples and 2K validation

TABLE 7

Semantic segmentation performance on ADE20K dataset with SemanticFPN [132] and UperNet [133] segmentors. The FLOPs are computed over the whole segmentor at the resolution of $2048 \times 512 \times 3$. \ddagger means testing model with multi scale and flip augmentations.

Semantic Segmentation on ADE20K						
Method	Backbone	FLOPs	#Param	mIoU	mIoU \ddagger	
SemanticFPN	PVT-S [4]	225G	28M	42.0	42.0	
	PVTv2-B2 [21]	230G	29M	45.2	-	
	RegionViT-S+ [17]	236G	35M	45.3	-	
	DAT-T [32]	198G	32M	42.6	44.2	
	DAT-T++	212G	28M	48.4	48.8	
	PVT-M [4]	315G	48M	42.9	43.3	
	RegionViT-B+ [17]	459G	77M	47.5	-	
	PVTv2-B4 [21]	427G	66M	47.9	-	
	DAT-S [32]	320G	53M	46.1	48.5	
	DAT-S++	339G	57M	49.9	50.7	
UperNet	PVT-L [4]	420G	65M	43.5	43.9	
	PVTv2-B5 [21]	486G	86M	48.7	-	
	DAT-B [32]	481G	92M	47.0	49.0	
	DAT-B++	508G	97M	50.4	51.1	
	Swin-T [3]	945G	60M	44.5	45.8	
	ConvNeXt-T [98]	939G	60M	46.1	46.7	
	Focal-T [16]	998G	62M	45.8	47.0	
	InternImage-T [103]	944G	59M	47.9	48.1	
	NAT-T [96]	934G	58M	47.1	48.4	
	DiNAT-T [97]	934G	58M	47.8	48.8	
Swin-B [3]	HorNet-T [101]	924G	55M	49.2	49.3	
	DAT-T [32]	957G	60M	45.5	46.4	
	DAT-T++	969G	54M	49.4	50.3	
	NAT-S [96]	1038G	81M	47.6	49.5	
	ConvNeXt-S [98]	1010G	82M	48.0	49.5	
	DiNAT-S [97]	1027G	82M	48.7	49.6	
	Focal-S [16]	1010G	82M	48.9	49.9	
	HorNet-S [101]	1130G	85M	48.0	50.0	
	InternImage-S [103]	1027G	85M	50.0	50.5	
	DAT-S [32]	1017G	80M	50.1	50.9	
ConvNeXt-B [98]	DAT-S++	1079G	81M	48.3	49.8	
	Swin-B [3]	1098G	85M	50.5	51.2	
	NAT-B [96]	1098G	121M	48.1	49.7	
	ConvNeXt-B [98]	1137G	123M	48.5	49.7	
	DiNAT-B [97]	1170G	122M	49.1	49.9	
	Focal-B [16]	1137G	123M	49.6	50.4	
	HorNet-B [101]	1354G	126M	49.0	50.5	
	InternImage-B [103]	1171G	126M	50.5	50.9	
	DAT-B [32]	1185G	128M	50.8	51.3	
	DAT-B++	1212G	121M	49.4	50.6	

examples. We employ our DAT on two widely adopted segmentation models, SemanticFPN [132] and UperNet [133]. Both frameworks share an encoder-decoder structure where we employ variants of DAT++ as encoder networks. We follow the recipes in Swin Transformer [3] and PVT [4] to initialize the encoder from ImageNet pretrained weights, then train 160K iterations for UperNet and 80K for SemanticFPN. We train DAT++ on the top of SemanticFPN [132] with 0.4, 0.4, 0.7 drop path rates for *Tiny*, *Small*, and *Base*

variants. And for UperNet [133], these factors are 0.3, 0.5, 0.7 for DAT-T/S/B++, respectively.

Results. Table 7 presents the results of different variants of DAT++ with two segmentation frameworks on ADE20K. The first part lists the results with SemanticFPN, from which DAT-T/S/B++ achieves 48.4, 49.9, and 50.4 in single-scale mIoU, 51.1 in multi-scale mIoU, respectively, which leads to sharp performance boosts over other methods. The similar superiority of DAT++ holds for UperNet, as shown in the second part. Taking ConvNeXt [101] as an example, DAT++ has a consistent improvement over HorNet on each of the three model scales, with +3.6, +2.6 and +1.6 in multi-scale mIoU, respectively. In addition, DAT++ outperforms many baselines and performs similar to SOTA methods, such as HorNet [101] and InternImage [103].

4.4 Ablation Study

To better understand our model, we study the key components in the proposed DMHA module and some important design choices of DAT and DAT++ individually, including offset learning, deformable positional encoding, local modeling, and macro designs. We first sketch a roadmap from the original DAT [32] to DAT++ with *Tiny* variant, displaying the steps toward a more powerful model with clean and effective modifications. We then select a version of DAT-T++ without augmented convolutions, denoted as DAT-T+, to ablate the design choices. At the end of this section, we provide a detailed analysis on the augmented convolutions to reveal how DAT-T+ is transformed into DAT-T++.

Roadmap from DAT to DAT++. As shown in Figure 5, we begin by removing the designs in the original DAT-T, which require too many extra hyper-parameters to build a concise architecture. Reducing the number of deformed keys and values from 14^2 to 7^2 in the 3rd stage makes it unified to incorporate DMHA modules in other stages with limited computational complexity, with 7^2 sampling points in all four stages. Another dispensable design is the restriction of the range of sampling offsets. The ablation in [32] makes it clear that a wide range from 1.0 to 10.0 of this restriction factor works well in many scenarios. Thus we remove it and just clip the sampling coordinates to ensure that these locations lie in the image. The above modifications bring about a -0.3 decline in accuracy but lead to a cleaner design.

To achieve better performance, we adopt several amendments in macro design. The deeper and narrower structure has been shown to be effective in many previous studies [7], [134], and therefore we increase the number of blocks from 6 to 18 in stage 3 of DAT while decreasing the dimensions of the basic model from 96 to 64 to alleviate the growth of complexity. In addition to depth and width, Neighborhood Attention [96] provides an efficient implementation of local attention with higher accuracy-speed performance than sliding window ones [87], [135]. We replace Window Attention [3] with Neighborhood Attention as the local attention module in DAT++. In addition, we extend the DMHA to the first two stages to build a full deformable attention model and further add two blocks in stage 2, leading to an accuracy of 83.0 with 4.05 GFLOPs. We denote this version as **DAT+** to ablate the design choices.

TABLE 8

Ablation study on the core components and designs in DMHA. **P** represents replacing DMHA with SRA attention proposed in [4], while **D** denotes the proposed DMHA attention module. ✓ in offsets means using learnable offsets to guide deformable feature sampling whereas ✗ means not. The FLOPs are computed at the resolution of 224^2 of the IN-1K classification model. Object detection is done with 1x Mask R-CNN and semantic segmentation is done with 160K UperNet.

Components	Backbone	IN-1K MS-COCO ADE20K					
Attn. Off.	Pos. Enc.	FLOPs #Param	Acc@1	AP ^b	AP ^m	mIoU	
P	✗	4.20G 25.8M	81.9	45.6	41.3	47.2	
D	✗	4.05G 22.6M	82.4	45.7	41.4	46.5	
D	✓	4.05G 22.6M	82.8	45.8	41.6	47.5	
D	✓	Fixed	4.05G 24.0M	82.8	45.8	41.5	47.8
D	✓	LogCPB [41]	4.13G 22.7M	82.5	46.3	41.6	47.8
D	✓	DeformRPB	4.05G 22.8M	83.0	46.2	41.8	47.9

TABLE 9

Ablation study on different types of local modeling operators. These modules play roles of **Local Attention** in Figure 3. FLOPs and parameters are computed at the resolution of 224^2 of image classification model. Our proposed **DAT+** is marked with gray line.

Local Operators	Backbone	IN-1K MS-COCO ADE20K			
	FLOPs #Param	Acc@1	AP ^b	AP ^m	mIoU
Window Attn. [3]	4.05G 22.8M	82.4	44.9	41.0	46.8
Slide Attn. [89]	3.57G 20.9M	82.5	45.9	41.6	47.7
Depth-wise Conv. [98]	3.39G 20.4M	82.3	46.1	41.8	47.7
Neighborhood Attn. [96]	4.05G 22.8M	83.0	46.2	41.8	47.9

To further push the limit, we enhance DAT+ with three convolutional modules introduced in Sec. 3.3.2 and end up with the final DAT++ with 83.9 accuracy and 4.29 GFLOPs, which achieves state-of-the-art performance. The convolutional components are analyzed as follows.

Core components and designs in DMHA. We study the core components and designs in the proposed DMHA that include learning offset and exploiting geometric information to demonstrate the effectiveness of DAT. We summarize these experiments in Table 8, in which the first part shows the cases of disabling offsets, and the second part shows the influence of different types of position encoding methods. DAT-T+ is shown in the last row, with 83.0 ImageNet-1K Top-1 accuracy and 46.2, 41.8 bbox and mask mAP on MS-COCO and 47.9 mIoU on ADE20K. To study the effect of offset learning, we come up with two variants of DAT+ without learning any sampling offsets for deformed keys and values by replacing the DMHA module with the SRA module in PVT [4] or simply zeroing out the offsets to sample features at the reference points. These two variants are shown in the first two rows, from which the sharp drops of up to 1.1 accuracy, 0.6 mAP and 1.4 mIoU demonstrate the importance of learning offsets and focusing on important image parts in the proposed DMHA module.

To study the most suitable way of exploiting geometric information for DMHA, we compare the performance of different position encoding methods in the third to fifth rows in Table 8. Enabling offset learning without using any position encoding brings about +0.4% accuracy, +0.1 mAP and +0.3 mIoU on different tasks. However, it makes hardly any difference to simply involve a fixed position bias in the deformable attention since there are no improvements on

TABLE 10

Ablation study on applying DMHA on different stages. Configuration **ND** denotes this stage is composed of successive Neighborhood Attention and DMHA blocks, as **NN** and **DD** denote the stage is made up of the same blocks of Neighborhood Attention and DMHA, respectively.

Stage Configurations				Backbone	IN-1K MS-COCO ADE20K			
1	2	3	4	FLOPs #Param	Acc@1	AP ^b	AP ^m	mIoU
NN	NN	NN	NN	4.29G 22.7M	81.4	43.9	39.7	45.3
NN	NN	NN	ND	4.29G 22.7M	81.9	45.0	40.9	45.5
NN	NN	NN	DD	4.29G 22.7M	82.3	45.2	41.2	46.7
NN	NN	ND	DD	4.21G 22.7M	82.8	46.4	41.8	47.5
NN	ND	ND	DD	4.08G 22.7M	83.0	46.3	41.7	47.7
ND	ND	ND	DD	4.05G 22.8M	83.0	46.2	41.8	47.9

TABLE 11

Ablation study on convolutional enhancements. ConvFFN means inserting a residual depth-wise convolution layer in MLP blocks, ConvStem means using overlapped patch embedding and downsampling modules between stages, LPU [86] means adding a residual depth-wise convolution layer before each attention block.

Added Convolution Modules	Backbone	IN-1K MS-COCO ADE20K			
	FLOPs #Param	Acc@1	AP ^b	AP ^m	mIoU
None (DAT-T+)	4.05G 22.8M	83.0	46.2	41.8	47.9
+ ConvFFN	4.12G 23.0M	83.4	47.5	42.6	48.6
+ ConvStem	4.27G 23.9M	83.6	48.2	43.2	49.0
+ LPU (DAT-T++)	4.29G 24.0M	83.9	48.7	43.7	49.4

image classification and object detection tasks, except for more parameters. We also try a light version of the LogCPB proposed in [41], by reducing the hidden dimensions of the MLP from 512 to 32 in order to maintain a feasible computation complexity and memory footprint. LogCPB exhibits good performance on object detection tasks with higher input resolutions, but image classification accuracy has a degradation of 0.3%. In addition, as the deformable attention works globally on the feature maps, LogCPB has larger FLOPs than other and scales poorly to high-resolution inputs. The last row reports the results of our proposed deformable relative positional bias, showing strong overall performance on various tasks, indicating that DeformRPB is the best suitable approach to exploiting geometric information in the scheme of deformable attention.

Locality modeling operators. We next investigate the combination of the proposed DMHA module with different operations to model the locality in the images. We replace the Neighborhood Attention [96] module with three other kinds of local operators: non-overlapped window attention in Swin Transformer [3], 7×7 convolution in ConvNeXt [98], and a novel local attention named Slide Attention that leverages depth-wise convolution [89] to achieve efficient sliding window attention. These modules play the role of **Local Attention** in Figure 3, whose performances on various vision tasks are listed in Table 9. When using non-overlapped window attention, the model only gets 82.4% ImageNet accuracy, even falling behind the one with 82.6% in Figure 5 whose first two stages are identical to Swin Transformer. In addition, the window attention variant also behaves poorly on downstream tasks by -0.2 descent in object detection and semantic segmentation. The novel Slide

Attention works much better than window attention with thorough improvements. Even though this variant has fewer parameters and less FLOPs, it has biquadratic memory footprints w.r.t. the kernel size, making it viable only in 3×3 kernels. Other than local attention methods, depth-wise convolutions can model locality efficiently, and many prior researchers have taken a step forward in modernizing CNNs with large kernels [98], [99], [100]. We choose the carefully designed 7×7 depth-wise convolution with MLP module in ConvNeXt to see its performance. The convolutional operator achieves similar results on MS-COCO object detection and ADE20K semantic segmentation with much fewer FLOPs, whereas it fails to close the gap in ImageNet classification. Therefore, we select Neighborhood Attention as the local attention module in DAT++. These results also imply the versatility of DAT with various local operators.

DMHA at different stages. We explore this design choice by gradually replacing Neighborhood Attention blocks with DMHA blocks stage by stage, as shown in Table 10. In the first row, all stages of the model are composed of Neighborhood Attention, which resembles NAT-T [96] in the block configurations. This version only has 81.4% accuracy in image classification, 43.9 mAP in object detection, and 45.3 mIoU in semantic segmentation. We begin by replacing one Neighborhood Attention block in the 4th stage with the proposed DMHA module, which immediately brings +0.5, +1.1, +1.2 and +0.2 gains on four metrics. Next we equip the whole 4th stage with DMHA, following as additional +0.4, +0.2, +0.3, +1.2 improvements. Keeping the procedure of DMHA replacements in the 3rd stage, more increments are obtained in these tasks. These sharp increments demonstrate the effectiveness of our proposed deformable attention. Although the increasing model performance as incorporating DMHA in more early stages is prone to saturate, we choose the final version with all four stages with deformable attention as DAT-T+ to construct as simple a model as possible.

Convolutional enhancements. Finally, we analyze how the convolutional enhancements in Sec. 3.3.2 can further boost the model performance and examine their contribution. Table 11 shows continuous improvements are made using ConvFFN, overlapped patch embeddings and downsamplings (+ConvStem), and local perception units (+LPU), step by step. We end up with DAT++ with 83.9% ImageNet-1K accuracy, 48.7 MS-COCO mAP, and 49.4 ADE20K mIoU, which is competitive with SOTA ViT/CNNs.

4.5 DAT and Deformable DETR

In this section, we compare two different types of deformable attention in DAT and in Deformable-DETR [23], which is a direct adaptation of deformable convolution [13]. We discuss the differences as follows:

The role of deformable attention. In DAT, deformable attention works in the backbone while serving as an efficient detection head in Deformable DETR to improve the convergence rate of the original DETR [106].

Attention mechanism. The m -th head of query q in single scale attention of Deformable DETR is formulated as

$$z_q^{(m)} = \sum_{k=1}^K A_{qk}^{(m)} W_v \phi(x; p_q + \Delta p_{qk}^{(m)}), \quad (16)$$

TABLE 12

Comparison on deformable attention in DAT and in Deformable-DETR under different computational budgets. The GPU memory cost is measured in a forward pass with 64 batch size under inference mode.

Deformable Attn. Types	#Keys	FLOPs	#Param	Peak Memory	IN-1K Top-1 Acc.
Deformable DETR [23]	16	4.17G	22.1M	11.5GB	79.2
	49	4.84G	25.6M	20.1GB	80.6
	100	5.89G	31.2M	33.6GB	81.1
DAT+	49	4.05G	22.8M	9.1GB	83.0

where K keys are sampled from the input features, mapped by W_v and then aggregated by attention weights $A_{qk}^{(m)}$. Compared to our Deformable Multi-Head Attention (DMHA, Eq. (11)), these attention weights are learned from z_q by a linear projection, i.e. $A_{qk}^{(m)} = \sigma(W_{\text{att}}x)$, where $W_{\text{att}} \in \mathbb{R}^{C \times MK}$ is a weight matrix to predict the weights for each key head by head, after which a softmax function σ is applied to normalize the attention score of K keys to $[0, 1]$. In fact, the attention weights are directly predicted by queries instead of comparing the similarities between queries and keys in the dot product manner. If we change the σ function to a sigmoid, it will be a variant of a modulated deformable convolution [104], resembling a dynamic convolution rather than attention.

Memory consumption. The deformable attention in Deformable DETR is not compatible with the dot product similarity computation for its huge memory footprints mentioned in Sec. 3.2.2. Therefore, the linear predicted attention with linear space complexity is used to avoid computing dot products, along with a smaller number of keys $K = 4$ to reduce the memory cost.

Empirical results. To validate our claim, we replace the DMHA in DAT with the modules in [23] to show that a naive adaptation is inferior for the vision backbone. We use DAT-T+ as a baseline, where the results are shown in Table 12. We vary the number of keys of the Deformable DETR model in 16, 49, and 100. The model $K = 16$ has FLOPs and parameters to DAT+, only achieving 79.2% accuracy. When using the same number of keys in 49, the memory cost becomes over 2x of DAT+, but the performance does not catch up. As the number of keys increases to 100, the memory, FLOPs and parameters all grow drastically. However, the accuracy of the Deformable DETR attention still has a gap of 1.9%.

4.6 Visualization

We visualize some qualitative results of DAT++ to verify the effectiveness of the proposed deformable attention, as shown in Figure 6. We present five examples from MS-COCO [30] validation dataset, using Mask R-CNN [130] (3x) equipped with DAT-T++ to perform instance segmentation. The first columns show the input images and the results output from the model.

The 3rd column of Figure 6 displays the important keys in the image, whose scores are represented by the sizes of the orange circles, where a larger circle indicates a higher score. For a comprehensive analysis of the importance score, we cumulatively multiply the attention weights of the deformed keys from the last DMHA layer to previous layers

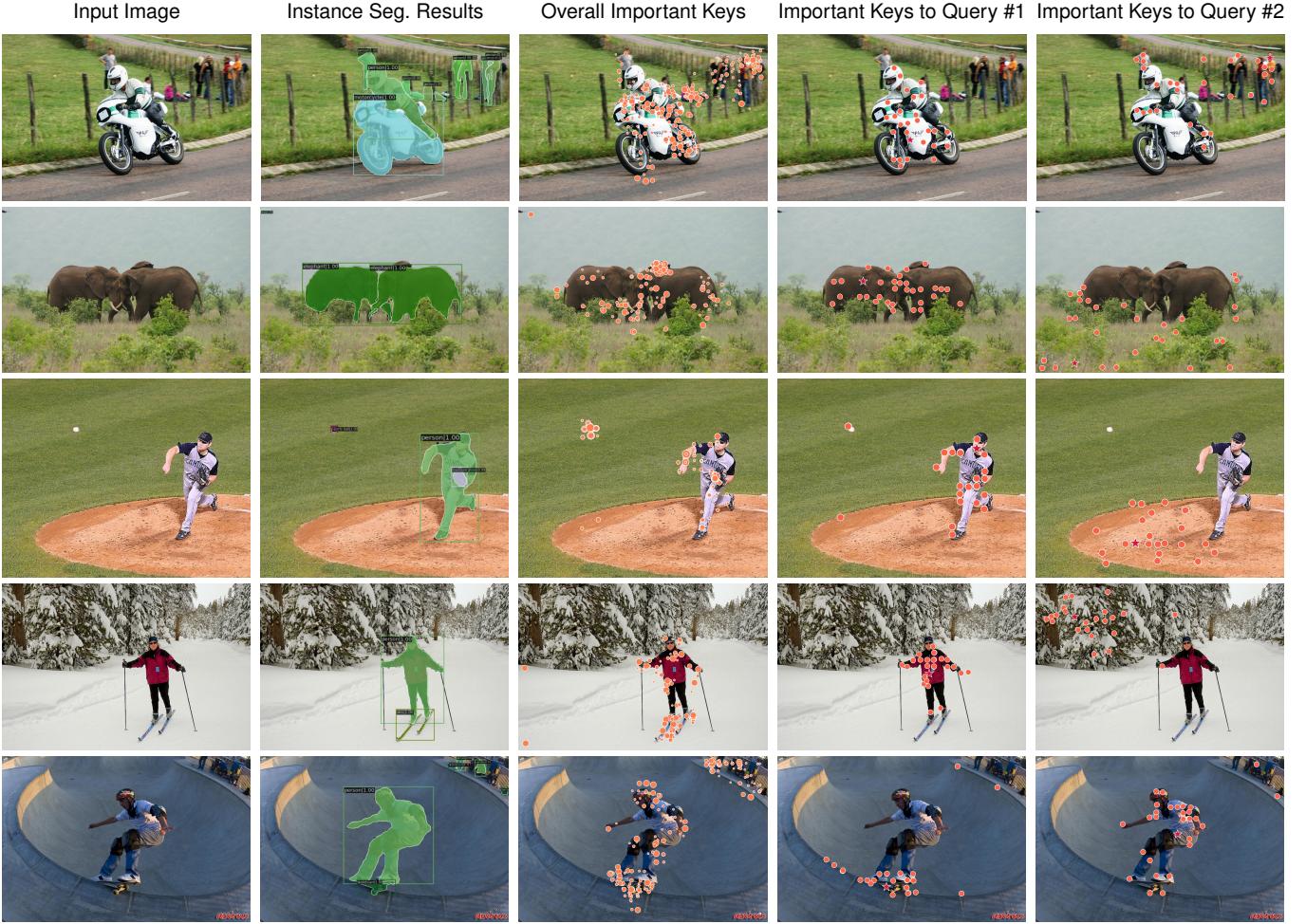


Fig. 6. Visualization of DAT++ with Mask R-CNN [130] on MS-COCO [30] validation set. From left to right, each column displays the input image, the bounding boxes and masks of instance segmentation, the overall important keys of each attention head, and the important keys to two given query points, respectively. In the 3rd column, each orange circle denotes a deformed key, whose larger radius reflects higher attention score. In the 4th and 5th columns, the query point is marked as a red star, and the keys with high attention scores are marked in coral. **Zoom in for best view.**

and then average them among all queries to discover the keys with the most contributions. We observe that the proposed deformable attention learns to attend the important keys mainly in the foreground, *e.g.*, in the 1st row, the keys with Top-10 accumulated attention scores are mostly distributed in the foreground motorbike and persons. Similar patterns are also found in the scene of the elephants, the baseball and pitcher, the skier and skis, and the skater and boards in the 2nd to 5th rows, indicating that the deformable attention can focus on the important regions of the objects, which supports our hypothesis illustrated in Figure 1.

In addition to the overall distribution of the deformed keys, we also provide visualization of the attention maps *w.r.t.* some specific query tokens, depicted in the last two columns of Figure 6. Taking the ski image in the 4th as an example, we place the query #1 (red star) on the skier, and the keys with high attention scores (coral circles) are concentrated mainly on the body of the skier. In comparison, for query #2 in the pine tree surround shown in the 5th column, its corresponding deformed keys scatter among the background trees, further demonstrating that DAT++ focuses on the keys that are shifted to the relevant parts of the queries with meaningful offsets.

5 CONCLUSION

This paper presents Deformable Attention Transformer, a novel hierarchical Vision Transformer that can be adapted to various visual recognition tasks, including image classification, object detection, instance segmentation, and semantic segmentation. With the proposed deformable attention and DMHA module, DAT and DAT++ are capable of learning spatially dynamic attention patterns in a data-dependent way and modeling geometric transformations. Extensive experiments demonstrate the effectiveness of DAT++ over other competitive ViT/CNNs. We hope our work can inspire insights towards designing dynamic visual attention mechanisms and more powerful foundation models.

ACKNOWLEDGMENTS

This work is supported in part by the National Key R&D Program of China under Grant 2021ZD0140407, in part by the National Natural Science Foundation of China under Grants 62022048 and 62276150, and in part by the Guoqiang Institute of Tsinghua University. We also appreciate the donation of computational resources from Hangzhou High-Flyer AI Fundamental Research Co.,Ltd.

REFERENCES

- [1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *NeurIPS*, 2017. [1](#)
- [2] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *ICLR*, 2020. [1, 2, 3, 5, 6](#)
- [3] 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 *ICCV*, 2021. [1, 2, 3, 5, 6, 7, 8, 9, 10, 11, 12](#)
- [4] 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 *ICCV*, 2021. [1, 2, 3, 6, 8, 10, 11, 12](#)
- [5] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016. [1](#)
- [6] G. Huang, Z. Liu, G. Pleiss, L. Van Der Maaten, and K. Weinberger, "Convolutional networks with dense connectivity," *IEEE TPAMI*, 2019. [1](#)
- [7] I. Radosavovic, R. P. Kosaraju, R. Girshick, K. He, and P. Dollár, "Designing network design spaces," in *CVPR*, 2020. [1, 11](#)
- [8] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017. [1](#)
- [9] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *CVPR*, 2018. [1](#)
- [10] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan *et al.*, "Searching for mobilenetv3," in *ICCV*, 2019. [1](#)
- [11] G. Huang, S. Liu, L. Van der Maaten, and K. Q. Weinberger, "Condensenet: An efficient densenet using learned group convolutions," in *CVPR*, 2018. [1](#)
- [12] L. Yang, H. Jiang, R. Cai, Y. Wang, S. Song, G. Huang, and Q. Tian, "Condensenet v2: Sparse feature reactivation for deep networks," in *CVPR*, 2021. [1](#)
- [13] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in *ICCV*, 2017. [1, 2, 3, 4, 13](#)
- [14] P. Zhang, X. Dai, J. Yang, B. Xiao, L. Yuan, L. Zhang, and J. Gao, "Multi-scale vision longformer: A new vision transformer for high-resolution image encoding," *arXiv preprint arXiv:2103.15358*, 2021. [1, 2, 8](#)
- [15] X. Dong, J. Bao, D. Chen, W. Zhang, N. Yu, L. Yuan, D. Chen, and B. Guo, "Cswin transformer: A general vision transformer backbone with cross-shaped windows," in *CVPR*, 2022. [1, 2, 8, 9, 10](#)
- [16] J. Yang, C. Li, P. Zhang, X. Dai, B. Xiao, L. Yuan, and J. Gao, "Focal attention for long-range interactions in vision transformers," in *NeurIPS*, 2021. [1, 2, 8, 9, 11](#)
- [17] C.-F. Chen, R. Panda, and Q. Fan, "Regionvit: Regional-to-local attention for vision transformers," in *ICLR*, 2022. [1, 2, 8, 9, 11](#)
- [18] Y. Yuan, R. Fu, L. Huang, W. Lin, C. Zhang, X. Chen, and J. Wang, "Hrformer: High-resolution vision transformer for dense predict," in *NeurIPS*, 2021. [1, 2](#)
- [19] H. Fan, B. Xiong, K. Mangalam, Y. Li, Z. Yan, J. Malik, and C. Feichtenhofer, "Multiscale vision transformers," in *ICCV*, 2021. [1, 2](#)
- [20] Y. Li, C.-Y. Wu, H. Fan, K. Mangalam, B. Xiong, J. Malik, and C. Feichtenhofer, "Mvitv2: Improved multiscale vision transformers for classification and detection," in *CVPR*, 2022. [1, 2](#)
- [21] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pvt v2: Improved baselines with pyramid vision transformer," *Computational Visual Media*, 2022. [1, 7, 8, 9, 11](#)
- [22] W. Zeng, S. Jin, W. Liu, C. Qian, P. Luo, W. Ouyang, and X. Wang, "Not all tokens are equal: Human-centric visual analysis via token clustering transformer," in *CVPR*, 2022. [1, 3](#)
- [23] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable detr: Deformable transformers for end-to-end object detection," in *ICLR*, 2021. [2, 3, 4, 13](#)
- [24] Z. Chen, Y. Zhu, C. Zhao, G. Hu, W. Zeng, J. Wang, and M. Tang, "Dpt: Deformable patch-based transformer for visual recognition," in *ACM MM*, 2021. [2, 3, 6, 8](#)
- [25] X. Yue, S. Sun, Z. Kuang, M. Wei, P. Torr, W. Zhang, and D. Lin, "Vision transformer with progressive sampling," in *ICCV*, 2021. [2, 3](#)
- [26] Y. Cao, J. Xu, S. Lin, F. Wei, and H. Hu, "Gcnet: Non-local networks meet squeeze-excitation networks and beyond," in *ICCVW*, 2019. [2, 4](#)
- [27] D. Zhou, B. Kang, X. Jin, L. Yang, X. Lian, Z. Jiang, Q. Hou, and J. Feng, "Deepvit: Towards deeper vision transformer," *arXiv preprint arXiv:2103.11886*, 2021. [2, 4](#)
- [28] X. Ma, H. Wang, C. Qin, K. Li, X. Zhao, J. Fu, and Y. Fu, "A close look at spatial modeling: From attention to convolution," *arXiv preprint arXiv:2212.12552*, 2022. [2, 4](#)
- [29] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *CVPR*. Ieee, 2009, pp. 248–255. [2, 7, 9](#)
- [30] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *ECCV*. Springer, 2014, pp. 740–755. [2, 7, 10, 13, 14](#)
- [31] B. Zhou, H. Zhao, X. Puig, T. Xiao, S. Fidler, A. Barriuso, and A. Torralba, "Semantic understanding of scenes through the ade20k dataset," *IJCV*, vol. 127, no. 3, pp. 302–321, 2019. [2, 7, 10](#)
- [32] Z. Xia, X. Pan, S. Song, L. E. Li, and G. Huang, "Vision transformer with deformable attention," in *CVPR*, 2022. [2, 5, 6, 7, 8, 9, 10, 11](#)
- [33] J. Gu, H. Kwon, D. Wang, W. Ye, M. Li, Y.-H. Chen, L. Lai, V. Chandra, and D. Z. Pan, "Multi-scale high-resolution vision transformer for semantic segmentation," in *CVPR*, 2022. [2](#)
- [34] Y. Xu, Q. Zhang, J. Zhang, and D. Tao, "Vitae: Vision transformer advanced by exploring intrinsic inductive bias," in *NeurIPS*, 2021. [2](#)
- [35] R. Yang, H. Ma, J. Wu, Y. Tang, X. Xiao, M. Zheng, and X. Li, "Scalablevit: Rethinking the context-oriented generalization of vision transformer," in *ECCV*, 2022. [2, 7](#)
- [36] Z. Tu, H. Talebi, H. Zhang, F. Yang, P. Milanfar, A. Bovik, and Y. Li, "Maxvit: Multi-axis vision transformer," in *ECCV*, 2022. [2](#)
- [37] M. Ding, B. Xiao, N. Codella, P. Luo, J. Wang, and L. Yuan, "Davit: Dual attention vision transformers," in *ECCV*, 2022. [2](#)
- [38] W. Wang, W. Chen, Q. Qiu, L. Chen, B. Wu, B. Lin, X. He, and W. Liu, "Crossformer++: A versatile vision transformer hinging on cross-scale attention," *arXiv preprint arXiv:2303.06908*, 2023. [2](#)
- [39] H. Huang, X. Zhou, J. Cao, R. He, and T. Tan, "Vision transformer with super token sampling," in *CVPR*, 2023. [2, 3, 6](#)
- [40] L. Zhu, X. Wang, Z. Ke, W. Zhang, and R. Lau, "Biformer: Vision transformer with bi-level routing attention," in *CVPR*, 2023. [2, 3, 6, 8, 9](#)
- [41] Z. Liu, H. Hu, Y. Lin, Z. Yao, Z. Xie, Y. Wei, J. Ning, Y. Cao, Z. Zhang, L. Dong *et al.*, "Swin transformer v2: Scaling up capacity and resolution," in *CVPR*, 2022. [2, 5, 12](#)
- [42] S. Sun, X. Yue, S. Bai, and P. H. S. Torr, "Visual parser: Representing part-whole hierarchies with transformers," *arXiv preprint arXiv:2107.05790*, 2021. [2](#)
- [43] A. Jaegle, F. Gimeno, A. Brock, O. Vinyals, A. Zisserman, and J. Carreira, "Perceiver: General perception with iterative attention," in *ICML*, 2021. [2](#)
- [44] J. Fang, L. Xie, X. Wang, X. Zhang, W. Liu, and Q. Tian, "Msg-transformer: Exchanging local spatial information by manipulating messenger tokens," in *CVPR*, 2022. [2](#)
- [45] J. Yang, C. Li, X. Dai, and J. Gao, "Focal modulation networks," in *NeurIPS*, 2022. [2](#)
- [46] J. Min, Y. Zhao, C. Luo, and M. Cho, "Peripheral vision transformer," in *NeurIPS*, 2022. [2](#)
- [47] Q. Zhang, Y. Xu, J. Zhang, and D. Tao, "Vitaev2: Vision transformer advanced by exploring inductive bias for image recognition and beyond," *International Journal of Computer Vision*, pp. 1–22, 2023. [2](#)
- [48] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang *et al.*, "Deep high-resolution representation learning for visual recognition," *IEEE TPAMI*, vol. 43, no. 10, pp. 3349–3364, 2020. [2](#)
- [49] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jegou, "Training data-efficient image transformers & distillation through attention," in *ICML*, vol. 139, July 2021, pp. 10 347–10 357. [2, 3, 6, 8](#)
- [50] H. Touvron, M. Cord, A. Sablayrolles, G. Synnaeve, and H. Jegou, "Going deeper with image transformers," in *ICCV*, 2021. [2, 6, 8](#)

- [51] K. Wu, J. Zhang, H. Peng, M. Liu, B. Xiao, J. Fu, and L. Yuan, "Tinyvit: Fast pretraining distillation for small vision transformers," in *ECCV*, 2022.
- [52] Z. Jiang, Q. Hou, L. Yuan, D. Zhou, X. Jin, A. Wang, and J. Feng, "Token labeling: Training a 85.5% top-1 accuracy vision transformer with 56m parameters on imagenet," *arXiv preprint arXiv:2104.10858*, 2021. [2](#)
- [53] H. Touvron, M. Cord, A. El-Nouby, J. Verbeek, and H. Jégou, "Three things everyone should know about vision transformers," in *ECCV*, 2022. [2](#)
- [54] H. Touvron, M. Cord, and H. Jégou, "Deit iii: Revenge of the vit," in *ECCV*, 2022. [2](#)
- [55] A. Steiner, A. Kolesnikov, X. Zhai, R. Wightman, J. Uszkoreit, and L. Beyer, "How to train your vit? data, augmentation, and regularization in vision transformers," *arXiv preprint arXiv:2106.10270*, 2021. [2, 3](#)
- [56] X. Chen, S. Xie, and K. He, "An empirical study of training self-supervised vision transformers," in *ICCV*, 2021. [2](#)
- [57] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *ICCV*, 2021. [2](#)
- [58] M. Oquab, T. Dariseti, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby *et al.*, "Dinov2: Learning robust visual features without supervision," *arXiv preprint arXiv:2304.07193*, 2023. [2](#)
- [59] Y. Li, H. Fan, R. Hu, C. Feichtenhofer, and K. He, "Scaling language-image pre-training via masking," in *CVPR*, 2023. [2](#)
- [60] Z. Xie, Z. Zhang, Y. Cao, Y. Lin, J. Bao, Z. Yao, Q. Dai, and H. Hu, "Simmim: A simple framework for masked image modeling," in *CVPR*, 2022. [2](#)
- [61] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *CVPR*, 2022. [2](#)
- [62] H. Bao, L. Dong, S. Piao, and F. Wei, "Beit: Bert pre-training of image transformers," in *ICLR*, 2021. [2](#)
- [63] J. Zhou, C. Wei, H. Wang, W. Shen, C. Xie, A. Yuille, and T. Kong, "Image BERT pre-training with online tokenizer," in *ICLR*, 2022. [2](#)
- [64] C. Wei, H. Fan, S. Xie, C.-Y. Wu, A. Yuille, and C. Feichtenhofer, "Masked feature prediction for self-supervised visual pre-training," in *CVPR*, 2022. [2](#)
- [65] W. Su, X. Zhu, C. Tao, L. Lu, B. Li, G. Huang, Y. Qiao, X. Wang, J. Zhou, and J. Dai, "Towards all-in-one pre-training via maximizing multi-modal mutual information," in *CVPR*, 2023. [2](#)
- [66] Y. Fang, W. Wang, B. Xie, Q. Sun, L. Wu, X. Wang, T. Huang, X. Wang, and Y. Cao, "Eva: Exploring the limits of masked visual representation learning at scale," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 19 358–19 369. [2](#)
- [67] M. Assran, Q. Duval, I. Misra, P. Bojanowski, P. Vincent, M. Rabbat, Y. LeCun, and N. Ballas, "Self-supervised learning from images with a joint-embedding predictive architecture," in *CVPR*, 2023. [2](#)
- [68] M. Dehghani, J. Djolonga, B. Mustafa, P. Padlewski, J. Heek, J. Gilmer, A. P. Steiner, M. Caron, R. Geirhos, I. Alabdulmohsin *et al.*, "Scaling vision transformers to 22 billion parameters," in *ICML*, 2023. [2](#)
- [69] M. Wortsman, G. Ilharco, S. Y. Gadre, R. Roelofs, R. Gontijo-Lopes, A. S. Morcos, H. Namkoong, A. Farhadi, Y. Carmon, S. Kormlith *et al.*, "Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time," in *ICML*, 2022. [2, 3](#)
- [70] Y. Han, G. Huang, S. Song, L. Yang, H. Wang, and Y. Wang, "Dynamic neural networks: A survey," *IEEE TPAMI*, vol. 44, no. 11, pp. 7436–7456, 2021. [2](#)
- [71] Y. Wang, R. Huang, S. Song, Z. Huang, and G. Huang, "Not all images are worth 16x16 words: Dynamic transformers for efficient image recognition," in *NeurIPS*, 2021. [2](#)
- [72] Y. Rao, Z. Liu, W. Zhao, J. Zhou, and J. Lu, "Dynamic spatial sparsification for efficient vision transformers and convolutional neural networks," *IEEE TPAMI*, 2023. [2](#)
- [73] Y. Xu, Z. Zhang, M. Zhang, K. Sheng, K. Li, W. Dong, L. Zhang, C. Xu, and X. Sun, "Evo-vit: Slow-fast token evolution for dynamic vision transformer," in *AAAI*, 2022. [2](#)
- [74] L. Meng, H. Li, B.-C. Chen, S. Lan, Z. Wu, Y.-G. Jiang, and S.-N. Lim, "Adavit: Adaptive vision transformers for efficient image recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 12 309–12 318. [2](#)
- [75] H. Yin, A. Vahdat, J. Alvarez, A. Mallya, J. Kautz, and P. Molchanov, "A-ViT: Adaptive tokens for efficient vision transformer," in *CVPR*, 2022. [2](#)
- [76] Y. Liang, C. GE, Z. Tong, Y. Song, J. Wang, and P. Xie, "EVit: Expediting vision transformers via token reorganizations," in *ICLR*, 2022. [2](#)
- [77] D. Bolya, C.-Y. Fu, X. Dai, P. Zhang, C. Feichtenhofer, and J. Hoffman, "Token merging: Your ViT but faster," in *ICLR*, 2023. [2](#)
- [78] S. Tang, J. Zhang, S. Zhu, and P. Tan, "Quadtree attention for vision transformers," in *ICLR*, 2022. [3](#)
- [79] M. Ding, Y. Shen, L. Fan, Z. Chen, Z. Chen, P. Luo, J. B. Tenenbaum, and C. Gan, "Visual dependency transformers: Dependency tree emerges from reversed attention," in *CVPR*, 2023. [3](#)
- [80] L. Song, S. Zhang, S. Liu, Z. Li, X. He, H. Sun, J. Sun, and N. Zheng, "Dynamic grained encoder for vision transformers," in *NeurIPS*, 2021. [3](#)
- [81] Y. Xie, J. Zhang, Y. Xia, A. v. d. Hengel, and Q. Wu, "Clustr: Exploring efficient self-attention via clustering for vision transformers," *arXiv preprint arXiv:2208.13138*, 2022. [3](#)
- [82] Q. Zhang, Y. Xu, J. Zhang, and D. Tao, "Vsa: Learning varied-size window attention in vision transformers," in *ECCV*, 2022. [3](#)
- [83] Q. Zhang, J. Zhang, Y. Xu, and D. Tao, "Vision transformer with quadrangle attention," *arXiv preprint arXiv:2303.15105*, 2023. [3](#)
- [84] W. Xu, Y. Xu, T. Chang, and Z. Tu, "Co-scale conv-attentional image transformers," *arXiv preprint arXiv:2104.06399*, 2021. [3](#)
- [85] Z. Dai, H. Liu, Q. V. Le, and M. Tan, "Coatnet: Marrying convolution and attention for all data sizes," in *NeurIPS*, 2021. [3](#)
- [86] J. Guo, K. Han, H. Wu, Y. Tang, X. Chen, Y. Wang, and C. Xu, "Cmt: Convolutional neural networks meet vision transformers," in *CVPR*, 2022. [3, 6, 7, 12](#)
- [87] H. Zhao, J. Jia, and V. Koltun, "Exploring self-attention for image recognition," in *CVPR*, 2020. [3, 11](#)
- [88] X. Pan, C. Ge, R. Lu, S. Song, G. Chen, Z. Huang, and G. Huang, "On the integration of self-attention and convolution," in *CVPR*, 2022. [3](#)
- [89] X. Pan, T. Ye, Z. Xia, S. Song, and G. Huang, "Slide-transformer: Hierarchical vision transformer with local self-attention," in *CVPR*, 2023. [3, 12](#)
- [90] S. d'Ascoli, H. Touvron, M. L. Leavitt, A. S. Morcos, G. Biroli, and L. Sagun, "Convit: Improving vision transformers with soft convolutional inductive biases," in *ICML*, 2021, pp. 2286–2296. [3](#)
- [91] H. Wu, B. Xiao, N. Codella, M. Liu, X. Dai, L. Yuan, and L. Zhang, "Cvt: Introducing convolutions to vision transformers," *arXiv preprint arXiv:2103.15808*, 2021. [3](#)
- [92] X. Chu, Z. Tian, B. Zhang, X. Wang, and C. Shen, "Conditional positional encodings for vision transformers," in *ICLR*, 2023. [3, 5, 6](#)
- [93] T. Xiao, M. Singh, E. Mintun, T. Darrell, P. Dollár, and R. Girshick, "Early convolutions help transformers see better," *arXiv preprint arXiv:2106.14881*, 2021. [3](#)
- [94] Q. Han, Z. Fan, Q. Dai, L. Sun, M.-M. Cheng, J. Liu, and J. Wang, "On the connection between local attention and dynamic depthwise convolution," in *ICLR*, 2022. [3](#)
- [95] M. Arar, A. Shamir, and A. H. Bermano, "Learned queries for efficient local attention," in *CVPR*, 2022. [3](#)
- [96] A. Hassani, S. Walton, J. Li, S. Li, and H. Shi, "Neighborhood attention transformer," in *CVPR*, 2023. [3, 7, 8, 9, 10, 11, 12, 13](#)
- [97] A. Hassani and H. Shi, "Dilated neighborhood attention transformer," *arXiv preprint arXiv:2209.15001*, 2022. [3, 7, 8, 9, 10, 11](#)
- [98] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," in *CVPR*, 2022. [3, 7, 8, 9, 10, 11, 12, 13](#)
- [99] X. Ding, X. Zhang, J. Han, and G. Ding, "Scaling up your kernels to 31x31: Revisiting large kernel design in cnns," in *CVPR*, 2022. [3, 7, 13](#)
- [100] S. Liu, T. Chen, X. Chen, X. Chen, Q. Xiao, B. Wu, T. Kärkkäinen, M. Pechenizkiy, D. C. Mocanu, and Z. Wang, "More convnets in the 2020s: Scaling up kernels beyond 51x51 using sparsity," in *ICLR*, 2023. [3, 13](#)
- [101] Y. Rao, W. Zhao, Y. Tang, J. Zhou, S. N. Lim, and J. Lu, "Horonet: Efficient high-order spatial interactions with recursive gated convolutions," in *NeurIPS*, 2022. [3, 8, 9, 10, 11](#)
- [102] Y. Rao, W. Zhao, Z. Zhu, J. Zhou, and J. Lu, "Gfnet: Global filter networks for visual recognition," *IEEE TPAMI*, 2023. [3](#)
- [103] W. Wang, J. Dai, Z. Chen, Z. Huang, Z. Li, X. Zhu, X. Hu, T. Lu, L. Lu, H. Li *et al.*, "Internimage: Exploring large-scale vision

- foundation models with deformable convolutions," in *CVPR*, 2022. 3, 6, 8, 9, 10, 11
- [104] X. Zhu, H. Hu, S. Lin, and J. Dai, "Deformable convnets v2: More deformable, better results," in *CVPR*, 2019. 3, 13
- [105] Z. Gao, Z. Tong, L. Wang, and M. Z. Shou, "Sparseformer: Sparse visual recognition via limited latent tokens," *arXiv preprint arXiv:2304.03768*, 2023. 3
- [106] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *ECCV*. Springer, 2020, pp. 213–229. 3, 13
- [107] G. Huang, Y. Wang, K. Lv, H. Jiang, W. Huang, P. Qi, and S. Song, "Glance and focus networks for dynamic visual recognition," *IEEE TPAMI*, 2022. 3
- [108] Y. Han, G. Huang, S. Song, L. Yang, Y. Zhang, and H. Jiang, "Spatially adaptive feature refinement for efficient inference," *IEEE TIP*, 2021. 3
- [109] Y. Pu, Y. Wang, Z. Xia, Y. Han, Y. Wang, W. Gan, Z. Wang, S. Song, and G. Huang, "Adaptive rotated convolution for rotated object detection," in *ICCV*, 2023. 3
- [110] Y. Han, D. Han, Z. Liu, Y. Wang, X. Pan, Y. Pu, C. Deng, J. Feng, S. Song, and G. Huang, "Dynamic perceiver for efficient visual recognition," in *ICCV*, 2023. 3
- [111] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016. 3, 6
- [112] A. Srinivas, T.-Y. Lin, N. Parmar, J. Shlens, P. Abbeel, and A. Vaswani, "Bottleneck transformers for visual recognition," in *CVPR*, 2021, pp. 16519–16529. 6
- [113] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *CVPR*, 2017, pp. 2117–2125. 6
- [114] Z. Pan, J. Cai, and B. Zhuang, "Fast vision transformers with hilo attention," in *NeurIPS*, 2022. 7
- [115] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *ICML*, 2015. 7
- [116] X. Chu, Z. Tian, Y. Wang, B. Zhang, H. Ren, X. Wei, H. Xia, and C. Shen, "Twins: Revisiting the design of spatial attention in vision transformers," in *NeurIPS*, 2021. 7
- [117] X. Pan, Z. Xia, S. Song, L. E. Li, and G. Huang, "3d object detection with pointformer," in *CVPR*, 2021. 7
- [118] B. Chen, P. Li, C. Li, B. Li, L. Bai, C. Lin, M. Sun, J. Yan, and W. Ouyang, "Glit: Neural architecture search for global and local image transformer," in *ICCV*, October 2021, pp. 12–21. 7
- [119] K. Han, A. Xiao, E. Wu, J. Guo, C. Xu, and Y. Wang, "Transformer in transformer," *arXiv preprint arXiv:2103.00112*, 2021. 7
- [120] A. Hatamizadeh, H. Yin, G. Heinrich, J. Kautz, and P. Molchanov, "Global context vision transformers," in *ICML*, 2023. 7
- [121] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *ICCV*, 2017, pp. 2980–2988. 8, 10
- [122] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017. 8
- [123] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le, "RandAugment: Practical automated data augmentation with a reduced search space," in *CVPRW*, 2020, pp. 702–703. 8
- [124] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, 2017. 8
- [125] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, "Cutmix: Regularization strategy to train strong classifiers with localizable features," in *ICCV*, 2019, pp. 6023–6032. 8
- [126] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," in *AAAI*, 2020. 8
- [127] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *JMLR*, 2014. 8
- [128] G. Huang, Y. Sun, Z. Liu, D. Sedra, and K. Q. Weinberger, "Deep networks with stochastic depth," in *ECCV*. Springer, 2016, pp. 646–661. 8
- [129] B. T. Polyak and A. B. Juditsky, "Acceleration of stochastic approximation by averaging," *SIAM journal on control and optimization*, vol. 30, no. 4, pp. 838–855, 1992. 8
- [130] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *ICCV*, 2017, pp. 2961–2969. 9, 10, 13, 14
- [131] Z. Cai and N. Vasconcelos, "Cascade r-cnn: Delving into high quality object detection," in *CVPR*, 2018, pp. 6154–6162. 10
- [132] A. Kirillov, R. Girshick, K. He, and P. Dollár, "Panoptic feature pyramid networks," in *CVPR*, 2019, pp. 6399–6408. 11
- [133] T. Xiao, Y. Liu, B. Zhou, Y. Jiang, and J. Sun, "Unified perceptual parsing for scene understanding," in *ECCV*, 2018. 11
- [134] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *ICML*, 2019. 11
- [135] P. Ramachandran, N. Parmar, A. Vaswani, I. Bello, A. Levskaya, and J. Shlens, "Stand-alone self-attention in vision models," in *NeurIPS*, 2019. 11