LocalViT: Bringing Locality to Vision Transformers

Yawei Li[1] Kai Zhang[1] Jiezhang Cao[1] Radu Timofte[1] Luc Van Gool[1,2]
[1]Computer Vision Lab, ETH Zurich, Switzerland [2]KU Leuven, Belgium
{yawei.li, kai.zhang, jiezhang.cao, timofter, vangool}@vision.ee.ethz.ch

Abstract

We study how to introduce locality mechanisms into vision transformers. The transformer network originates from machine translation and is particularly good at modelling long-range dependencies within a long sequence. Although the global interaction between the token embeddings could be well modelled by the self-attention mechanism of transformers, what is lacking a locality mechanism for information exchange within a local region. Yet, locality is essential for images since it pertains to structures like lines, edges, shapes, and even objects.

We add locality to vision transformers by introducing depth-wise convolution into the feed-forward network. This seemingly simple solution is inspired by the comparison between feed-forward networks and inverted residual blocks. The importance of locality mechanisms is validated in two ways: 1) A wide range of design choices (activation function, layer placement, expansion ratio) are available for incorporating locality mechanisms and all proper choices can lead to a performance gain over the baseline, and 2) The same locality mechanism is successfully applied to 4 vision transformers, which shows the generalization of the locality concept. In particular, for ImageNet2012 classification, the locality-enhanced transformers outperform the baselines DeiT-T [41] and PVT-T [46] by 2.6% and 3.1% with a negligible increase in the number of parameters and computational effort. Code is available at https://github.com/ofsoundof/LocalViT.

1. Introduction

Convolutional neural networks (CNNs) now define the state-of-the-art for computer vision tasks such as image classification [22, 35, 16, 20], object detection [11, 33], segmentation [27, 15], low-level vision [7, 53], *etc.* CNNs are based on locality in that convolutional filters only perceive a local region of the input image, *i.e.* the receptive field. By stacking multiple layers, the effective receptive fields of a deep neural network can be enlarged progressively. This design enables the network to learn a hierarchy of deep features, which is essential for the success of CNNs. Meanwhile, the local, repetitive connections save many parameters compared with fully connected layers. Yet, one problem is that a larger receptive field can only be achieved by combining layers, despite alternative attempts at enlarging the receptive field [51].

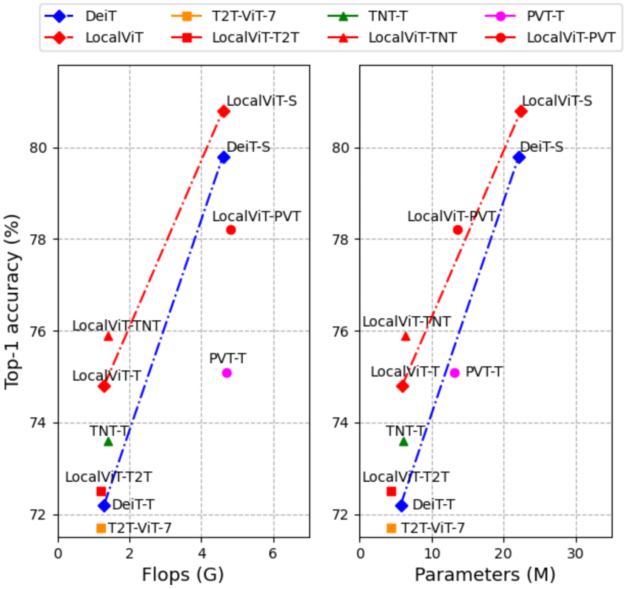


Figure 1: Comparison between LocalViT and the baseline transformers. The transformers enhanced by the proposed locality mechanism outperform their baselines.

A parallel, thriving research strand incorporates global connectivity into the network via self-attention [42, 43, 28, 48]. This family of networks, *i.e.* transformer networks, originates from machine translation and is very good at modelling long-range dependencies in sequences. There also is a rising interest in applying transformers to vision [2, 8, 41]. Vision transformers have already achieved performances quite competitive with their CNN counterparts.

To process 2D images with transformers, the input image is first converted to a sequence of tokens which correspond to patches in the image. Then the attention module attends to all tokens and a weighted sum is computed as the tokens for the next layer. In this way, the effective receptive field is expanded to the whole image via a single self-attention layer. Yet, the problem of visual transformers is that global connectivity contradicts the convolutional idea.

Considering the merits of CNNs vs. transformers, a natural question is *whether we can efficiently combine the locality of CNNs and the global connectivity of vision transformers to improve performance while not increasing model complexity.*

We try to fill the gap between CNNs and vision transformers. Specifically, we introduce a locality mechanism to the feed-forward network of transformers, which is inspired by examining the feed-forward network and inverted residuals [34, 17]. The feed-forward network of transformers consists of two fully connected layers and the hidden dimension between them is expanded (usually by a factor of 4) to extract richer features. Similarly, in inverted residual blocks, the hidden channel between the two 1×1 convolutions is also expanded. The major difference between them is the efficient depth-wise convolution in the inverted residual block. Such depth-wise convolution can provide precisely the mechanism for local information aggregation which is missing in the feed-forward network of vision transformers. In addition, depth-wise convolution is efficient in both parameters and computational complexity.

To cope with the 2D depth-wise convolution, the image tokens of the sequence from the self-attention module must be rearranged to a 2D feature map, which is processed by the feed-forward network. The class token is split out and bypasses the feed-forward network. The derived new feature map is converted back to image tokens and concatenated with the bypassed class token. The concatenated sequence is processed by the next transformer layer.

The effectiveness of the introduced locality mechanism is validated in two ways. Firstly, its properties are investigated experimentally. We draw four basic conclusions. *i.* Depth-wise convolution alone can already improve the performance of the baseline transformer. *ii.* A better activation function after depth-wise convolution can result in a significant performance gain. *iii.* The locality mechanism is more important for lower layers. *iv.* Expanding the hidden dimension of the feed-forward network leads to a larger model capacity and a higher classification accuracy. Secondly, as shown in Fig. 1, the locality mechanism is successfully applied to 4 vision transformers, which underlines its generality. The contributions of this paper are three-fold:

1. We bring a locality mechanism to vision transformers by introducing depth-wise convolutions. The new transformer architecture combines a self-attention mechanism for global relationship modelling and a locality mechanism for local information aggregation.
2. We analyze the basic properties of the introduced locality mechanism. The influence of each component (depth-wise convolution, non-linear activation function, layer placement, and hidden dimension expansion ratio) is singled out.
3. We apply these ideas to vision transformers incl. DeiT [41], T2T-ViT [52], PVT [46], and TNT [14]. Experiments show that the simple technique proposed in this paper generalizes well to various transformer architectures.

2. Related Work

2.1. Transformers and vision transformers

Transformers were first introduced in [42] for machine translation. The proposed attention mechanism aggregates information from the whole input sequence. Thus, transformers are especially good at modelling long-range dependencies between elements of a sequence. Since then, there have been several attempts to adapt transformers towards vision tasks including object detection [2, 56], image classification [8, 41, 52, 46, 14], segmentation [44], multiple object tracking [37, 29], human pose estimation [50, 55], point cloud processing [12, 54], video processing [10, 31, 38], image super-resolution [30, 49, 3], image synthesis [9], etc. An extensive review is out of the scope of this paper. We focus on the most relevant works.

Carion *et al.* first proposed a detection transformer (DETR) for end-to-end objection detection [2]. This method regards object detection as a set prediction problem and removes the hand-crafted designs for objection detection. DETR reasons about the relationship between the learned object queries and global image context. Following this work, image classification was targeted. Dosovitskiy *et al.* showed that a pure transformer can be directly applied to images and performs quite well compared with CNNs on image classification [8]. Yet, this network relies heavily on large-scale models and datasets. Thus, Touvron *et al.* showed that it is possible to train vision transformers in a data-efficient way [41]. The authors introduced an additional distillation token to the network and proposed hard-label distillation for vision transformers. Such transformers are identical to those for machine translation. Recent works propose to adapt transformers to images. Yuan *et al.* proposed a progressive tokenization method that can model the local information of nearby tokens and reduce the number of tokens. Wang *et al.* propose a pyramid architecture for vision transformers [46]. Han *et al.* introduced an additional transformer block for the image token embeddings [14].

2.2. Locality *vs.* global connectivity

Both local information and global connectivity help to reason about the relationships between image contents.

Thus, they are both important for visual perception. The convolution operation applies a sliding window to the input and local information is inherently aggregated to compute new representations. Thus, locality is an intrinsic property of CNNs [23]. Although CNNs can extract information from a larger receptive field by stacking layers and forming deep networks, they still lack global connectivity [22, 35, 16]. To overcome this problem, some researchers add global connectivity to CNNs with non-local blocks [47, 25].

By contrast, transformers are especially good at modelling long-range dependencies within a a sequence owing to their attention mechanism [42, 6, 4, 21]. But, in return, a locality mechanism remains to be added for visual perception. Some works already contributed towards this goal [52, 26]. Yet, they mainly focus on improving the tokenization and self-attention part. Other work introduces hybrid architectures of CNNs and transformers [2, 36, 24]. In summary, little attention has been paid to the feed-forward network of vision transformers.

2.3. Depth-wise convolution and inverted residuals

Compared with normal convolution, the computations of depth-wise convolution are only conducted channel-wise. That is, to obtain a channel of the output feature map, the convolution is only conducted on one input feature map. Thus, depth-wise convolution is efficient both in terms of parameters and computation. Thus, Howard *et al.* first proposed the MobileNet architecture based on depth-wise separable convolutions [18]. This lightweight and computationally efficient network is quite friendly for mobile devices. Since then, depth-wise convolution has been widely used to design efficient models. Inverted residual blocks are based on depth-wise convolution and were first introduced in MobileNetV2 [34]. The inverted residual blocks are composed of a sequence of 1×1 - depth-wise -1×1 convolutions. The hidden dimension between the two 1×1 convolutions is expanded. The utilization of depth-wise convolution avoids the drastic increase of model complexity brought by normal convolution. Due to the efficiency of this module, it is widely used to form the search space of neural architecture search (NAS) [17, 39, 40, 13, 1]. The expansion of the hidden dimension of inverted residuals is quite similar to the feed-forward network of vision transformers. This motivates us to think about the connection between them (See Sec. 3.2).

3. Methodology

Transformers are usually composed of encoders and decoders with similar building blocks. For the image classification task considered here, only the encoders are included in the network. Thus, we mainly describe the operations in the encoder layers. The encoders have two components, *i.e.* the self-attention mechanism that relates a token to all of the tokens and a feed-forward network that is applied to every token. We specifically explain how to introduce locality into the feed-forward network.

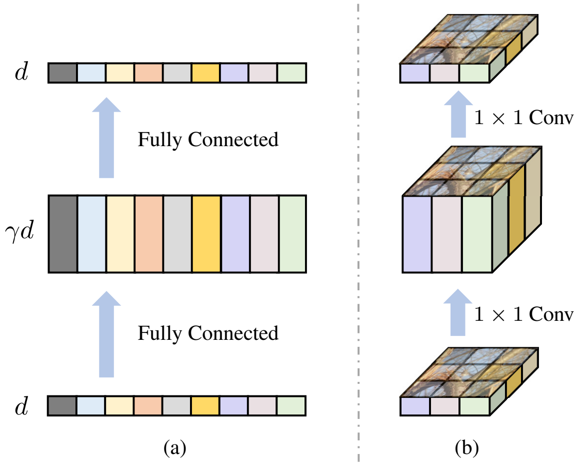


Figure 2: Visualization of the feed-forward network in transformers from different perspectives. (a) The input is regarded as a sequence of tokens. (b) An equivalent perspective is to still rearrange the tokens as a 2D lattice.

3.1. Input interpretation

Sequence perspective. Inherited from language modelling, transformers regard the input as a sequence that contains elements of embedded vectors. Consider an input image $\mathbf{X} \in \mathbb{R}^{C \times H \times W}$, where C and $H \times W$ denote the channel and spatial dimension of the input image, respectively. The input image is first converted to a sequence of tokens $\{\hat{\mathbf{X}}_i \in \mathbb{R}^d | i = 1, \ldots, N\}$, where $d = C \times p^2$ is the embedding dimension and $N = \frac{HW}{p^2}$. The tokens can be aggregated into a matrix $\hat{\mathbf{X}} \in \mathbb{R}^{N \times d}$.

Self-attention. In the self-attention mechanism, the relationship between the tokens is modelled by the similarity between the projected query-key pairs, yielding the attention score. The new tokens are computed as the weighted sum of the project values. That is,

$$\mathbf{Z} = \text{Softmax}(\mathbf{Q}\mathbf{K}^T / \sqrt{d})\mathbf{V}, \quad (1)$$

where the Softmax function is applied to the rows of the similarity matrix and d provides a normalization. The query, key, and value are a projection of the tokens, *i.e.* $\mathbf{Q} = \hat{\mathbf{X}}\mathbf{W}_Q$, $\mathbf{K} = \hat{\mathbf{X}}\mathbf{W}_K$, $\mathbf{V} = \hat{\mathbf{X}}\mathbf{W}_V$. The projection matrices $\mathbf{W}_Q$ and $\mathbf{W}_K$ have the same size while $\mathbf{W}_V$ could have a different size. In practice, the three projection matrices usually have the same size, *i.e.* $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V \in \mathbb{R}^{d \times d}$.

Feed-forward network. After the self-attention layer, a feed-forward network is appended. The feed-forward network consists of two fully-connected layers and transforms the features along the embedding dimension. The hidden dimension between the two fully-connected layers is expanded to learn a richer feature representation. That is,

$$\mathbf{Y} = f(\mathbf{Z}\mathbf{W}_1)\mathbf{W}_2, \tag{2}$$

where $\mathbf{W}_1 \in \mathbb{R}^{d\times\gamma d}$, $\mathbf{W}_2 \in \mathbb{R}^{\gamma d\times d}$, and $f(\cdot)$ denotes a non-linear activation function. For the sake of simplicity, the bias term is omitted. The dimension expansion ratio γ is usually set to 4. As shown in Fig. 2 (a), the input to the feed-forward network is regarded as a sequence of embedding vectors.

Lattice perspective. Since the feed-forward network is applied position-wise to a sequence of tokens $\mathbf{Z} \in \mathbb{R}^{N\times d}$, an exactly equivalent representation is to rearrange the sequence of tokens into a 2D lattice as shown in Fig. 2 (b). Then the reshaped feature representation is

$$\mathbf{Z}^r = \mathrm{Seq2Img}(\mathbf{Z}), \mathbf{Z}^r \in \mathbb{R}^{h\times w\times d}, \tag{3}$$

where $h = H/p$ and $w = W/p$. The operation Seq2Img converts a sequence to a 2D feature map. Each token is placed to a pixel location of the feature map. The benefit of this perspective is that the proximity between tokens is recovered, which provides the chance to introduce locality into the network. The fully-connected layers could be replaced by 1×1 convolutions, *i.e.*

$$\mathbf{Y}^r = f(\mathbf{Z}^r \circledast \mathbf{W}_1^r) \circledast \mathbf{W}_2^r, \tag{4}$$
$$\mathbf{Y} = \mathrm{Img2Seq}(\mathbf{Y}^r), \tag{5}$$

where $\mathbf{W}_1^r \in \mathbb{R}^{d\times\gamma d\times 1\times 1}$ and $\mathbf{W}_2^r \in \mathbb{R}^{\gamma d\times d\times 1\times 1}$ are reshaped from $\mathbf{W}_1$ and $\mathbf{W}_2$ and represent the convolutional kernels. The operation Img2Seq converts the image feature map back to a token sequence which is used by the next self-attention layer.

3.2. Locality

Since only 1×1 convolution is applied to the feature map, there is a lack of information interaction between adjacent pixels. Besides, the self-attention part of the transformer only captures global dependencies between all of the tokens. Thus, the transformer block does not have a mechanism to model the local dependencies between nearby pixels. It would be interesting if locality could be brought to transformers in an efficient way.

The expansion of the hidden dimension between fully-connected layers and the lattice perspective of the feed-forward network remind us of the inverted residual block proposed in MobileNets [34, 17]. As shown in Fig. 3, both of the feed-forward network and the inverted residual expand and squeeze the hidden dimension by 1×1 convolution. The only difference is that there is a depth-wise convolution in the inverted residual block. Depth-wise convolution applies a $k \times k$ ($k > 1$) convolution kernel per channel. The features inside the $k \times k$ kernel is aggregated to compute a new feature. Thus, depth-wise convolution is an efficient way of introducing locality into the network. Considering that, we reintroduce depth-wise convolution into the feed-forward network of transformers. And the computation could be represented as

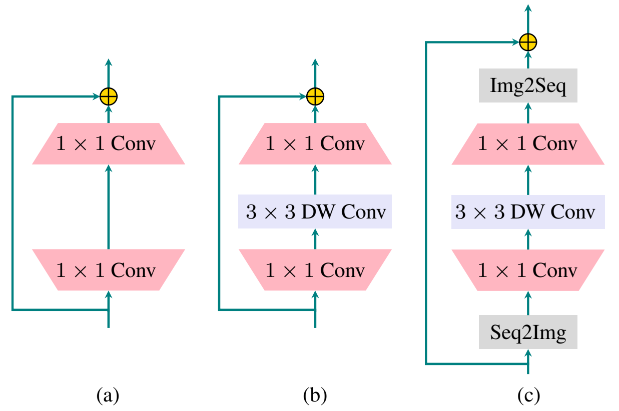


Figure 3: Comparison between the (a) convolutional version of the feed-forward network in vision transformers, the (b) inverted residual blocks, and (c) the finally utilized network that brings locality mechanism into transformers. "DW" denotes depth-wise convolution. To cope with the convolution operation, the conversion between sequence and image feature map is added by "Seq2Img" and "Img2Seq" in (c).

$$\mathbf{Y}^r = f\big(f(\mathbf{Z}^r \circledast \mathbf{W}_1^r) \circledast \mathbf{W}_d\big) \circledast \mathbf{W}_2^r, \tag{6}$$

where $\mathbf{W}_d \in \mathbb{R}^{\gamma d\times 1\times k\times k}$ is the kernel of the depth-wise convolution. The finally used network is shown in Fig. 3 (c). The input, *i.e.* a sequence of tokens is first reshaped to a feature map rearranged on a 2D lattice. Then two 1×1 convolutions along with a depth-wise convolution are applied to the feature map. After that, the feature map is reshaped to a sequence of tokens which are used as by the self-attention of the network transformer layer.

Note that the non-linear activation functions are not visualized in Fig. 3. Yet, they play a quite important role in enhancing the network capacity, especially for efficient networks. In particular, we try ReLU6, h-swish [17], squeeze-and-excitation (SE) module [19], efficient channel attention (ECA) module [45], and their combinations. A thorough analysis of the activation function is discussed in the experiments section.

3.3. Class token

To apply vision transformers to image classification, a trainable class token is added and inserted into the token embedding, *i.e.*

$$\hat{\mathbf{X}} \leftarrow \mathrm{Concat}(\mathbf{X}_{cls}, \hat{\mathbf{X}}), \tag{7}$$

Network	γ	DW	Params (M)	FLOPs (G)	Top-1 Acc. (%)
DeiT-T [41]	4	No	5.7	1.3	72.2
LocalViT-T	4	No	5.7	1.3	72.5 (0.3↑)
LocalViT-T*	4	Yes	5.8	1.3	73.7 (1.5↑)
DeiT-T [41]	6	No	7.5	1.6	73.1†
LocalViT-T	6	No	7.5	1.6	74.3 (1.2↑)
LocalViT-T*	6	Yes	7.7	1.6	76.1 (3.0↑)

Table 1: Investigation of the locality brought by depth-wise convolution. A comparison is made with DeiT. "DW" denotes depth-wise convolution. "Params" denotes the number of parameters in the network. "FLOPs" denotes the number of floating-point operations. Top-1 accuracy is reported. The other tables use the same evaluation metric. *ReLU6 is used as the activation function after depth-wise convolution. †Results derived by modifying the DeiT architecture and training the network with the same training protocol.

where $\leftarrow$ denotes the assignment operation, $\mathbf{X}_{cls} \in \mathbb{R}^{1\times d}$ is the class token. The new matrix has the dimension of $(N+1)\times d$ and $N+1 = \frac{HW}{p^2}+1$ tokens. In the self-attention module, the class token exchanges information with all other image tokens and gathers information for the final classification. In the feed-forward network, the same transformation is applied to the class token and the image tokens.

When depth-wise convolution is introduced into the feed-forward network, the sequence of tokens needs to be rearranged into an image feature map. Yet, the additional dimension brought by the class token makes the exact rearrangement impossible. To circumvent this problem, we split the $N+1$ tokens in Eqn. (1) into a class token and image tokens again, *i.e.*

$$(\mathbf{Z}_{cls}, \mathbf{Z}) \leftarrow \text{Split}(\mathbf{Z}). \tag{8}$$

Then the new image token is passed through the feed-forward network according to Eqns. (3), (6), and (5), leading to $\mathbf{Y}$. The class token is not passed through the feed-forward network. Instead, it is directly concatenated with $\mathbf{Y}$, *i.e.*

$$\mathbf{Y} \leftarrow \text{Concat}(\mathbf{Z}_{cls}, \mathbf{Y}). \tag{9}$$

The split and concatenation of the class token is done for every transformer layer. Although the class token $\mathbf{Z}_{cls}$ is not passed through the feed-forward network, the performance of the overall network is not adversely affected. This is because the information exchange and aggregation is done only in the self-attention part. A feed-forward network like Eqn. (2) only enforces a transformation within each token.

4. Experimental Results

This section gives the experimental results for image classification. We first study how the locality brought by depth-wise convolution can improve the performance of transformers. Then we show the influence of a non-linear activation function placed after the depth-wise convolution. The layers that are equipped with locality also influence the network capacity and this is also studied. An ablation study of the hidden dimension expansion ratio γ is also made. All those experiments are based on DeiT-T [41]. Finally, the study on the generalization to other vision transformers including T2T-ViT [52], PVT [46], TNT [14] for image classification and the comparison with CNNs are made. The transformers that are equipped with locality are denoted as LocalViT followed by the suffix that denotes the basic architecture.

4.1. Implementation details

In order to introduce locality into transformers, we only adapt the feed-forward network of vision transformers while the other parts such as self-attention, and position encoding are not changed. The implementation is based on the inverted residual blocks [34, 17]. A batch normalization layer is appended to the 2D convolutions. The layer normalization before the feed-forward network is removed. Path drop after the feed-forward network is also removed. The same modification is also applied to the T2T module of T2T-ViT [52]. The feed-forward network of the inner transformer block in TNT [14] is not adapted. For PVT [46], the class token is only introduced in the final stage of the pyramid. Thus, the split and concatenation of the class token for the feed-forward network is only applied in the final stage.

Experimental setup. The ImageNet2012 dataset [5] is used in this paper. The dataset contains 1.28M training images and 50K validation images from one thousand classes. We follow the same training protocol as DeiT [41]. The input image is randomly cropped with size 224 × 224. Cross-entropy is used as the loss function. Label smoothing is used. The weight decay factor is set to 0.05. The AdamW optimizer is used with a momentum of 0.9. The training continues for 300 epochs. The batch size is set to 1024. The initial learning rate is set to 1×10^{-3} and decreases to 1×10^{-5} following a cosine learning rate scheduler. During validation, a center crop of the validation images is conducted. In the following sections, we study different aspects that could influence the performance of the introduced locality.

4.2. Influence of the locality

We first study how the local information could help to improve the performance of vision transformers. Different hidden dimension expansion ratios γ are investigated.

Activation	Params (M)	FLOPs (G)	Top-1 Acc. (%)
Deit-T [41]	5.7	1.3	72.2
ReLU6	5.8	1.3	73.7 (1.5↑)
h-swish	5.8	1.3	74.4 (2.2↑)
h-swish + ECA	5.8	1.3	74.5 (2.3↑)
h-swish + SE-192	5.9	1.3	74.8 (2.6↑)
h-swish + SE-96	6.0	1.3	74.8 (2.6↑)
h-swish + SE-48	6.1	1.3	75.0 (2.8↑)
h-swish + SE-4	9.4	1.3	75.8 (3.6↑)

Table 2: Investigation of the non-linear activation function. The combination of h-swish, ECA [45], and SE [19] is studied. "SE-**" means the reduction ratio in the squeeze-and-excitation module. The study is based on LocalViT-T.

DW Placement	Layer	Params (M)	FLOPs (G)	Top-1 Acc. (%)
High	9∼12	5.78	1.26	69.1
Mid	5∼8	5.78	1.26	72.1
Low	1∼4	5.78	1.26	73.1
Low	1∼8	5.84	1.27	74.0
All	1∼12	5.91	1.28	74.8

Table 3: Influence of the placement of locality. "All" means all of the transformer layers are enhanced by depth-wise convolution. "Low", "Mid", and "High" mean the lower, middle, and higher transformer layers are equipped with depth-wise convolution, respectively. The study is based on LocalViT-T.

First of all, due to the change of the operations in the feed-forward network (Sec. 4.1), the Top-1 accuracy of LocalViT-T is slightly increased even without the depth-wise convolution. The performance gain is 0.3% for $\gamma = 4$ and is increased to 1.2% for $\gamma = 6$. Note that compared with DeiT-T, no additional parameters and computation are introduced for the improvement. When locality is incorporated into the feed-forward network, there is a significant improvement of the model accuracy, *i.e.* 1.5% for $\gamma = 4$ and 3.0% for $\gamma = 6$. Compared with the baseline, there only is a marginal increase in the number of parameters and a negligible increase in the amount of computation. **Thus, the performance of vision transformers can be significantly improved by the incorporation of a locality mechanism and the adaptation of the operation in the feed-forward network.**

4.3. Activation functions

The non-linear activation function after depth-wise convolution used in the above experiments is simply ReLU6. The benefit of using other non-linear activation functions is also studied. First of all, by replacing the activation function from ReLU6 to h-swish, the gain of Top-1 accuracy over the baseline is increased from 1.5% to 2.2%. This shows the benefit of h-swish activation functions can be easily extended from CNNs to vision transformers. Next, the h-swish activation function is combined with other channel attention modules including ECA [45] and SE [19]. By adding ECA, the performance is further improved by 0.1%. Considering that only 60 parameters are introduced, this improvement is still considerable under a harsh parameter budget.

Expansion Ratio γ	SE	Params (M)	FLOPs (G)	Top-1 Acc. (%)
1	No	3.1	0.7	65.9
	Yes	3.1	0.7	66.2
2	No	4.0	0.9	70.1
	Yes	4.0	0.9	70.6
3	No	4.9	1.1	72.9
	Yes	5.0	1.1	73.1
4	No	5.8	1.3	74.4
	Yes	5.9	1.3	74.8

Table 4: Investigating the expansion ratio of hidden layers in the feed-forward network.

Another significant improvement is brought by a squeeze-and-excitation module. When the reduction ratio in the squeeze-and-excitation module is reduced from 192 to 4, the gain of Top-1 accuracy is gradually increased from 2.6% to 3.6%. The number of parameters is also increased accordingly. Note that, for all of the networks, the computational complexity is almost the same. This implies that if there is no strict limitation on the number of parameters, advanced non-linear activation functions could be used. In the following experiments, we use the combination of h-swish and SE as the non-linear activation function after depth-wise convolution. Additionally, the reduction ratio of the squeeze-and-excitation module is chosen such that only 4 channels are kept after the squeeze operation. This choice of design achieves a good balance between the number of parameters and the model accuracy. **Thus, local information is also important in vision transformers. A wide range of efficient modules could be introduced into the feed-forward network of vision transformers to expand the network capacity.**

4.4. Placement of locality

The transformer layers where the locality is introduced can also influence the performance of the network. Thus, an ablation study based on LocalViT-T is conducted to study

Network	Image Size	Params (M)	FLOPs (G)	Top-1 Acc. (%)	Top-5 Acc. (%)
CNNs					
ResNet-18 [16]	224 × 224	11.7	1.8	69.8	89.1
ResNet-50 [16]	224 × 224	25.6	4.1	76.1	92.9
DenseNet-169 [20]	224 × 224	14.2	3.4	75.6	92.8
RegNet-4GF [32]	224 × 224	20.7	4.0	80.0	–
MobileNetV1 [18]	224 × 224	4.2	0.6	70.6	–
MobileNetV2 (1.4) [34]	224 × 224	6.9	0.6	74.7	–
EfficientNet-B0 [40]	224 × 224	5.3	0.4	77.1	93.3
EfficientNet-B4 [40]	380 × 380	19.3	4.5	82.9	96.4
Transformers					
DeiT-T [41]	224 × 224	5.7	1.3	72.2	91.1
LocalViT-T	224 × 224	5.9	1.3	74.8 (2.6↑)	92.6
DeiT-T⚗ [41]	224 × 224	5.9	1.3	74.5	–
DeiT-S [41]	224 × 224	22.1	4.6	79.8	95.1
LocalViT-S	224 × 224	22.4	4.6	80.8 (1.0↑)	95.4
DeiT-S⚗ [41]	224 × 224	22.4	4.6	81.2	–
T2T-ViT-7 [52]	224 × 224	4.3	1.2	71.7	–
LocalViT-T2T	224 × 224	4.3	1.2	72.5 (0.8↑)	–
TNT-T [14]	224 × 224	6.1	1.4	73.6	91.9
LocalViT-TNT	224 × 224	6.3	1.4	75.9 (2.3↑)	93.0
PVT-T [46]	224 × 224	13.2	4.7	75.1	92.3
LocalViT-PVT	224 × 224	13.5	4.8	78.2 (3.1↑)	94.2

Table 5: Image classification results for different CNNs and vision transformers. The locality functionality is enabled for four different vision transformers.

their effect. The results is reported in Table 3. There are in total 12 transformer layers in the network. We divide the 12 layers into 3 groups corresponding to "Low", "Mid", and "High" stages. For the former 3 rows of Table 3, we independently insert locality into the three stages. As the locality is moved gradually from lower stages to the higher stages, the accuracy of the network is decreased. This shows that local information is especially important for the lower layers. This is also consistent with our intuition. That is, when the depth-wise convolution is applied to the lower layers, the local information aggregated there could also be propagated to the higher layers, which is important to improve the overall performance of the network.

When the locality is introduced only in the higher stage, the Top-1 accuracy is even lower than DeiT-T. To investigate whether locality in the higher layers always has an adverse effect, we progressively allow more lower layers to have depth-wise convolution until locality is enabled for all layers. This corresponds to the last three rows of Table 3. The observation is that starting from the lower layers, the performance of the network could be gradually improved as locality is enabled for more layers. **Thus, introducing the locality to the lower layers is more advantageous compared with higher layers.**

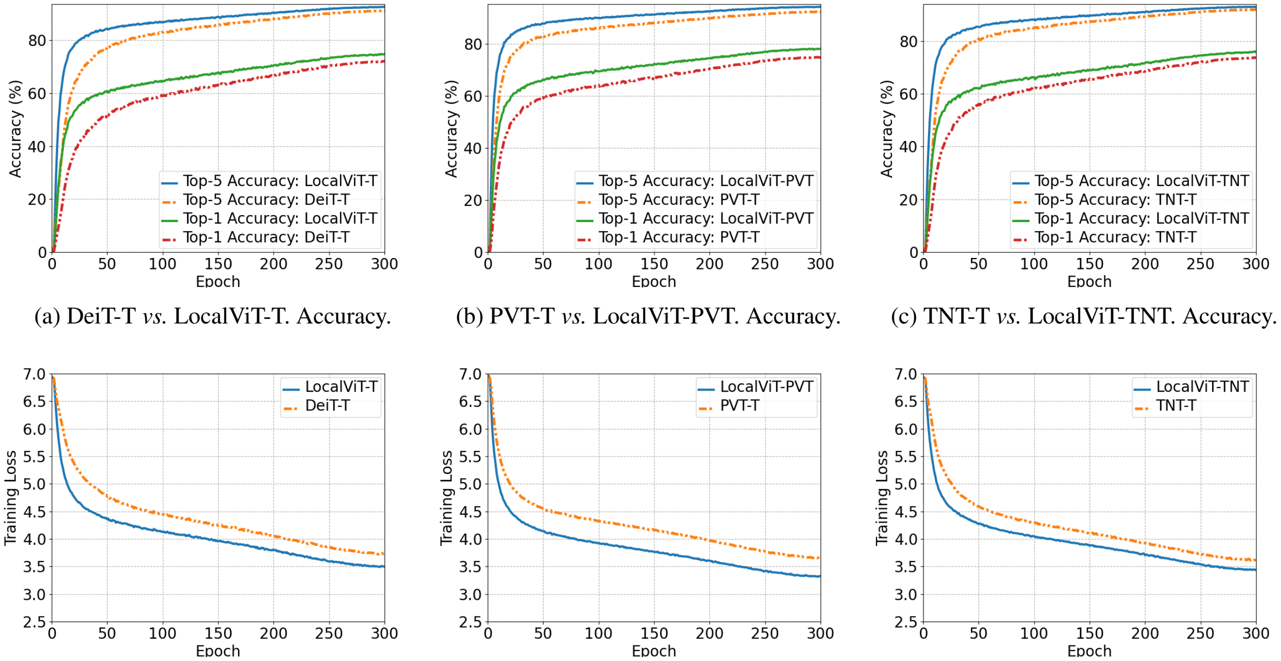
4.5. Expansion ratio γ

The effect of the expansion ratio of the hidden dimension of the feed-forward network is also investigated. The results are shown in Table 4. Expanding the hidden dimension of the feed-forward network can have a significant effect on the performance of the transformers. As the expansion ratio is increased from 1 to 4, the Top-1 accuracy is increased from less than 70% to nearly 75%. The model complexity is also almost doubled. **Thus, the network performance and model complexity can be balanced by the hidden dimension expansion ratio γ. Squeeze-and-excitation can be more beneficial for smaller γ.**

4.6. Generalization and comparison

Finally, we try to generalize the knowledge derived above to more vision transformers including DeiT-S [41], T2T-ViT [52], TNT [14], PVT [46] and compare their performance with CNNs. The result is shown in Table 5.

We draw two major conclusions from Table 5. **Firstly, the effectiveness of locality can be generalized to a wide range of vision transformers based on the following observations.** ***I.*** Compared with DeiT, LocalViT can yield a higher classification accuracy for both the tiny and small



(a) DeiT-T *vs.* LocalViT-T. Accuracy.

(b) PVT-T *vs.* LocalViT-PVT. Accuracy.

(c) TNT-T *vs.* LocalViT-TNT. Accuracy.

(d) DeiT-T *vs.* LocalViT-T. Training loss.

(e) PVT-T *vs.* LocalViT-PVT. Training loss.

(f) TNT-T *vs.* LocalViT-TNT. Training loss.

Figure 4: Comparison of Top-1 accuracy, Top-5 accuracy, and training loss between the baseline transformers and the locality enhanced LocalViT.

version of the network. The increase of Top-1 accuracy is 2.6% and 1.0%, resp. LocalViT-T even outperforms DeiT-T⚗ which is enhanced by knowledge distillation from RegNetY-160 [32]. The small version LocalViT-S is slightly worse than DeiT-S⚗ by 0.4%. ***II.*** LocalViT-T2T outperforms T2T-ViT-7 by 0.8%. Note that T2T-ViT already tries to model the local structure information in the tokens-to-token module. ***III.*** In TNT, an additional transformer block is used to extract local features for the image tokens. Thus, the locality is also considered in TNT. The modified network, *i.e.* LocalViT-TNT could still improve the classification accuracy by a large margin of 2.3%. ***IV.*** The biggest improvement comes from PVT. Introducing the locality module leads to a gain of 3.1% over PVT-T. ***V.*** The comparison of the training log between the baseline transformers and LocalViT is shown in Fig. 4. During the training phase, LocalViT consistently leads to a lower training loss and higher validation accuracy than the baseline transformers.

Secondly, some versions of the enhanced vision transformer LocalViT are already quite comparable or even outperform CNNs. This conclusion can be drawn by making the pairwise comparison, *i.e.* LocalViT-T *vs.* MobileNetV2 (1.4), LocalViT-S *vs.* ResNet-50, LocalViT-T2T *vs.* MobileNetV1, LocalViT-PVT *vs.* DenseNet-169 *etc.*

5. Conclusion

In this paper, we proposed to incorporate a locality mechanism into vision transformers. This is done by incorporating 2D depth-wise convolutions followed by a non-linear activation function into the feed-forward network of vision transformers. The idea is motivated by the comparison between the feed-forward network of transformers and the inverted residuals of MobileNets. In previous works, the input to the feed-forward network is a sequence of tokens embedding converted from an image. To cope with the locality mechanism, the sequence of tokens embedding is rearranged into a lattice as a 2D feature map, which is used as the input to the enhanced feed-forward network. To enable the rearrangement, the class token is split before the feed-forward network and concatenated with other image embeddings after the feed-forward network. A series of studies were made to investigate various factors (activation function, layer placement, and expansion ratio) that might influence of performance of the locality mechanism. The proposed locality mechanism is successfully applied to four different vision transformers, which validates its generality.

Acknowledgements: This work was partly supported by the ETH Zürich Fund (OK), a Huawei Technologies Oy (Finland) project, and an Amazon AWS grant.

References

[1] Han Cai, Ligeng Zhu, and Song Han. Proxylessnas: Direct neural architecture search on target task and hardware. *arXiv preprint arXiv:1812.00332*, 2018. 3

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020. 1, 2, 3

[3] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. *arXiv preprint arXiv:2012.00364*, 2020. 2

[4] Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, et al. Rethinking attention with performers. *arXiv preprint arXiv:2009.14794*, 2020. 3

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proc. CVPR*, pages 248–255. IEEE, 2009. 5

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 3

[7] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *Proc. ECCV*, pages 184–199. Springer, 2014. 1

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1, 2

[9] Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis. *arXiv preprint arXiv:2012.09841*, 2020. 2

[10] Rohit Girdhar, Joao Carreira, Carl Doersch, and Andrew Zisserman. Video action transformer network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 244–253, 2019. 2

[11] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proc. CVPR*, pages 580–587, 2014. 1

[12] Meng-Hao Guo, Jun-Xiong Cai, Zheng-Ning Liu, Tai-Jiang Mu, Ralph R Martin, and Shi-Min Hu. Pct: Point cloud transformer. *arXiv preprint arXiv:2012.09688*, 2020. 2

[13] Zichao Guo, Xiangyu Zhang, Haoyuan Mu, Wen Heng, Zechun Liu, Yichen Wei, and Jian Sun. Single path one-shot neural architecture search with uniform sampling. In *European Conference on Computer Vision*, pages 544–560. Springer, 2020. 3

[14] Kai Han, An Xiao, Enhua Wu, Jianyuan Guo, Chunjing Xu, and Yunhe Wang. Transformer in transformer. *arXiv preprint arXiv:2103.00112*, 2021. 2, 5, 7

[15] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 1

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. CVPR*, pages 770–778, 2016. 1, 3, 7

[17] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In *Proc. ICCV*, pages 1314–1324, 2019. 2, 3, 4, 5

[18] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. MobileNets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017. 3, 7

[19] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 4, 6

[20] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proc. CVPR*, pages 2261–2269, 2017. 1, 7

[21] Nikita Kitaev, Łukasz Kaiser, and Anselm Levskaya. Reformer: The efficient transformer. *arXiv preprint arXiv:2001.04451*, 2020. 3

[22] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Proc. NeurIPS*, pages 1097–1105, 2012. 1, 3

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

[24] Changlin Li, Tao Tang, Guangrun Wang, Jiefeng Peng, Bing Wang, Xiaodan Liang, and Xiaojun Chang. Bossnas: Exploring hybrid cnn-transformers with block-wisely self-supervised neural architecture search. *arXiv preprint arXiv:2103.12424*, 2021. 3

[25] Ding Liu, Bihan Wen, Yuchen Fan, Chen Change Loy, and Thomas S Huang. Non-local recurrent network for image restoration. *arXiv preprint arXiv:1806.02919*, 2018. 3

[26] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv preprint arXiv:2103.14030*, 2021. 3

[27] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proc. CVPR*, pages 3431–3440, 2015. 1

[28] Sachin Mehta, Marjan Ghazvininejad, Srinivasan Iyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. Delight: Very deep and light-weight transformer. *arXiv preprint arXiv:2008.00623*, 2020. 1

[29] Tim Meinhardt, Alexander Kirillov, Laura Leal-Taixe, and Christoph Feichtenhofer. Trackformer: Multi-object tracking with transformers. *arXiv preprint arXiv:2101.02702*, 2021. 2

[30] Chong Mou, Jian Zhang, Xiaopeng Fan, Hangfan Liu, and Ronggang Wang. Cola-net: Collaborative attention network for image restoration. *arXiv preprint arXiv:2103.05961*, 2021. 2

[31] Daniel Neimark, Omri Bar, Maya Zohar, and Dotan Asselmann. Video transformer network. *arXiv preprint arXiv:2102.00719*, 2021. 2

[32] Ilija Radosavovic, Raj Prateek Kosaraju, Ross Girshick, Kaiming He, and Piotr Dollár. Designing network design spaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10428–10436, 2020. 7, 8

[33] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Proc. NeurIPS*, pages 91–99, 2015. 1

[34] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. MobileNetV2: Inverted residuals and linear bottlenecks. In *Proc. CVPR*, pages 4510–4520, 2018. 2, 3, 4, 5, 7

[35] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1, 3

[36] Aravind Srinivas, Tsung-Yi Lin, Niki Parmar, Jonathon Shlens, Pieter Abbeel, and Ashish Vaswani. Bottleneck transformers for visual recognition. *arXiv preprint arXiv:2101.11605*, 2021. 3

[37] Peize Sun, Yi Jiang, Rufeng Zhang, Enze Xie, Jinkun Cao, Xinting Hu, Tao Kong, Zehuan Yuan, Changhu Wang, and Ping Luo. Transtrack: Multiple-object tracking with transformer. *arXiv preprint arXiv:2012.15460*, 2020. 2

[38] Jing Tan, Jiaqi Tang, Limin Wang, and Gangshan Wu. Relaxed transformer decoders for direct action proposal generation. *arXiv preprint arXiv:2102.01894*, 2021. 2

[39] Mingxing Tan, Bo Chen, Ruoming Pang, Vijay Vasudevan, Mark Sandler, Andrew Howard, and Quoc V Le. Mnasnet: Platform-aware neural architecture search for mobile. In *Proc. CVPR*, pages 2820–2828, 2019. 3

[40] Mingxing Tan and Quoc V Le. Efficientnet: Rethinking model scaling for convolutional neural networks. *arXiv preprint arXiv:1905.11946*, 2019. 3, 7

[41] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. *arXiv preprint arXiv:2012.12877*, 2020. 1, 2, 5, 6, 7

[42] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017. 1, 2, 3

[43] Hanrui Wang, Zhanghao Wu, Zhijian Liu, Han Cai, Ligeng Zhu, Chuang Gan, and Song Han. Hat: Hardware-aware transformers for efficient natural language processing. *arXiv preprint arXiv:2005.14187*, 2020. 1

[44] Huiyu Wang, Yukun Zhu, Bradley Green, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Axial-deeplab: Standalone axial-attention for panoptic segmentation. In *European Conference on Computer Vision*, pages 108–126. Springer, 2020. 2

[45] Qilong Wang, Banggu Wu, Pengfei Zhu, Peihua Li, Wangmeng Zuo, and Qinghua Hu. Eca-net: Efficient channel attention for deep convolutional neural networks. In *CVPR*, 2020. 4, 6

[46] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. *arXiv preprint arXiv:2102.12122*, 2021. 1, 2, 5, 7

[47] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018. 3

[48] Zhanghao Wu, Zhijian Liu, Ji Lin, Yujun Lin, and Song Han. Lite transformer with long-short range attention. *arXiv preprint arXiv:2004.11886*, 2020. 1

[49] Fuzhi Yang, Huan Yang, Jianlong Fu, Hongtao Lu, and Baining Guo. Learning texture transformer network for image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5791–5800, 2020. 2

[50] Sen Yang, Zhibin Quan, Mu Nie, and Wankou Yang. Transpose: Towards explainable human pose estimation by transformer. *arXiv preprint arXiv:2012.14214*, 2020. 2

[51] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015. 1

[52] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. *arXiv preprint arXiv:2101.11986*, 2021. 2, 3, 5, 7

[53] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a Gaussian denoiser: residual learning of deep CNN for image denoising. *IEEE TIP*, 26(7):3142–3155, 2017. 1

[54] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip Torr, and Vladlen Koltun. Point transformer. *arXiv preprint arXiv:2012.09164*, 2020. 2

[55] Ce Zheng, Sijie Zhu, Matias Mendieta, Taojiannan Yang, Chen Chen, and Zhengming Ding. 3d human pose estimation with spatial and temporal transformers. *arXiv preprint arXiv:2103.10455*, 2021. 2

[56] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 2