

Locally Enhanced Self-Attention: Combining Self-Attention and Convolution as Local and Context Terms

Chenglin Yang, Siyuan Qiao, Adam Kortylewski, Alan Yuille
Johns Hopkins University

{chenglin.yang, siyuan.qiao, akortyl1, ayuille1}@jhu.edu

Abstract

Self-Attention has become prevalent in computer vision models. Inspired by fully connected Conditional Random Fields (CRFs), we decompose self-attention into local and context terms. They correspond to the unary and binary terms in CRF and are implemented by attention mechanisms with projection matrices. We observe that the unary terms only make small contributions to the outputs, and meanwhile standard CNNs that rely solely on the unary terms achieve great performances on a variety of tasks. Therefore, we propose Locally Enhanced Self-Attention (LESA), which enhances the unary term by incorporating it with convolutions, and utilizes a fusion module to dynamically couple the unary and binary operations. In our experiments, we replace the self-attention modules with LESA. The results on ImageNet and COCO show the superiority of LESA over convolution and self-attention baselines for the tasks of image recognition, object detection, and instance segmentation. The code is made publicly available¹.

1. Introduction

Self-Attention has made a great influence in the computer vision community recently. It led to the emergence of fully attentional models [38, 51] and transformers [4, 17, 48]. Importantly, they show superior performances over traditional convolution neural networks on a variety of tasks including classification, object detection, segmentation, and image completion [31, 43, 49, 50].

Despite its remarkable achievement, the understanding of self-attention remains limited. One of its advantages is overcoming the limitation of spatial distances on dependency modelling. Originating from natural language processing, attention models the dependencies without regard to the distances among the words in the sequence, compared to LSTM [21] and gated RNN [13]. Being applied to vision models, attention aggregates the information globally

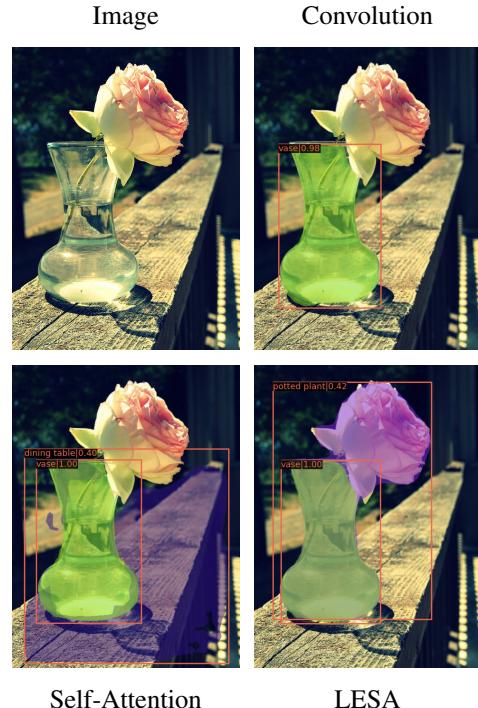


Figure 1. Effectiveness of Locally Enhanced Self-Attention (LESA) on COCO object detection and instance segmentation. All the outputs come from the Mask-RCNN models with WRN50 as the backbone. The training and testing processes are the same. The only difference is the spatial operations in the last two stages of WRN50. The comparisons are ensured fair. All the following visual examples are obtained by these models.

among the pixels or patches [17, 51]. Similarly, compared to traditional convolutions, the features extracted by attention are no longer constrained by a local neighborhood.

We argue that global aggregations in self-attention also bring problems because the aggregated features cannot clearly distinguish local and contextual cues. We study this from the perspective of Conditional Random Fields (CRFs) and decompose it into local and context terms. The unary

¹<https://github.com/Chenglin-Yang/LESA>

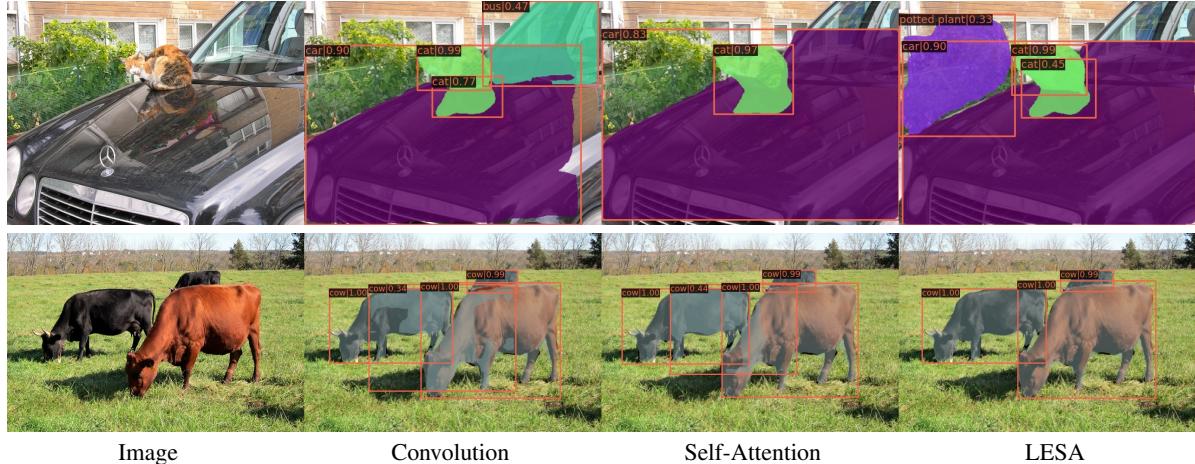


Figure 2. Effectiveness of LESA on COCO object detection and instance segmentation. For detection (first row), the existence of local terms enables LESA more capable of detecting objects in detail and correctly. For segmentation (second row), the combination of local and context terms makes LESA generate masks with higher semantic consistency.

(local) and binary (context) terms are based on the same building blocks of queries, keys, and values, and are calculated using the same projection matrices. We hypothesize that using the same building blocks for the local and context terms will cause problems, which relates to the weaknesses of projections in self-attention pointed out by Dong *et. al.* [16]. They theoretically prove that the output of consecutive self-attention layers converges doubly exponentially to a rank-1 matrix and verify this degeneration in transformers empirically. They also claim that skip connections can partially resolve the rank-collapse problem. In our CRF analysis, the skip connections create the simplest local term which amounts to the identity mapping. Skip connections alleviate the problem but we argue a local term with a stronger representation capacity needs to be designed.

In this paper, we enhance the unary term by integrating it with convolutions and propose Locally Enhanced Self-Attention (LESA), which is visualized in Fig. 3. To analyze self-attention from the perspective of CRFs, let x be the input and x^{+1} the output of one layer of self-attention. Both of them are two-dimensional grid of nodes. At spatial location (i, j) , the node $x_{i,j}^{+1}$ is connected to all the nodes $x_{h,w}$ of the input. The binary term involves the computation on the edges $x_{h \neq i, w \neq j} \rightarrow x_{i,j}^{+1}$ while the unary term the computation on the edge $x_{h=i, w=j} \rightarrow x_{i,j}^{+1}$. Intuitively, these two terms indicate the activation by looking at itself (local) and the others (context). Through ablation study in Tab. 1, we find that the unary term is important for the performance but only contributing to the output less than 2% computed by the softmax operation in attention. Without the unary term, the feature extraction at (i, j) entirely depends on interactions and losses the precise information of that pixel. The structure of the self-attention does not facil-

tate this unary operation. To address this issue, we enhance the unary term to $x_{(h,w) \in N(i,j)} \rightarrow x_{i,j}^{+1}$ where $N(i, j)$ indicate the pixels in neighborhood, and implement it as a grouped convolution followed by a projection layer.

To couple the unary and binary terms, we propose a dynamic fusion mechanism. The simplest static ways would be to assign equal weights to them or by setting their weights to be hyper-parameters. By contrast, we enable the model to allocate the weights on demand. Specifically, for each layer l with the binary term, we multiply the binary term element-wisely by $\omega^l(x) \in \mathbb{R}^{H \times W \times C}$. ω depends on the input x and dynamically controls the weights of the binary terms to the unary terms for different layers l , spatial locations H, W , and feature channels C .

We study the performance of LESA for image classification, object detection, and instance segmentation. We replace the spatial convolutions with LESAs in the last two stages of ResNet [19] and its larger variant WRN [58]. Then, we use them equipped with FPN [29] as the backbones in Mask-RCNN [18] to evaluate their performance for object detection and instance segmentation. The challenging large-scale datasets ILSVRC2012 [39] and COCO [30] are used to train and evaluate the models. The experiments demonstrate the superiority of LESA over the convolution and self-attention baselines.

To summarize, the main contributions of this work are:

- Analyzing self-attention from the perspective of fully connected CRFs, we decompose it into a pair of local (unary) and context (binary) terms. We observe the unary terms make small contributions to the outputs. Inspired by the standard CNNs’ focus on the local cues, we propose to enhance the unary term by incorporating it with convolutions.

- We propose a dynamic fusion module to couple the unary and binary terms adaptively. Their relative weights are adjusted as needed, depending on specific inputs, spatial locations, and feature channels.
- We implement Locally Enhanced Self-Attention (LESA) for vision tasks. Our experiments on the challenging datasets, ImageNet and COCO, demonstrate that the LESA is superior to the convolution and self-attention baselines. Especially for object detection and instance segmentation where local features are particularly important, LESA achieves significant improvements.

2. Related Work

2.1. Convolution

Convolutional Neural Networks (CNNs) have become the dominant models in computer vision in the last decade. AlexNet [27] shows considerable improvement over the models based on hand-crafted features [35, 40], and opens the door to the age of deep neural networks. Lots of efforts have been made to increase the width and depth and to improve the architecture and efficiency of CNNs in the pursuit of performance. They include the designs of VGG [42], GoogleNet [45], ResNet [19], WRN [58], ResNeXt [56], DenseNet [25], SENet [24], MobileNet [22], EfficientNet [46], etc. Through this process, the convolution layers are also being developed, leading to the grouped convolutions [56], depth-wise separable convolutions [11], deformable convolutions [14, 60], atrous convolutions [7, 33] and switchable atrous convolutions [9, 37].

2.2. Self-Attention

The impact of self-attention on vision community is becoming greater. Self-attention is originally proposed in approaches of natural machine translation [1]. It enables the encoder-decoder model to adaptively find the useful information according to contents from a variable length sentence. In computer vision, non-local neural networks [53] show that self-attention is an instantiation of non-local means [3], and use it to capture long-range dependencies to augment CNNs for tasks including video classification and object detection. A^2 -Net [10] employs a variant of non-local means and shows performance improvement on image classification. Recently, fully attentional methods [23, 38, 59] which replaces all the spatial convolutions with self-attention in the deep networks are proposed with stronger performances than CNNs. Axial attention [51] factorizes the 2D self-attention into two 1D consecutive self-attentions which reduces the computation complexity and enables the self-attention layer to have a global kernel. Self-attention also promotes the generation of transformers [4, 17, 31, 47, 48, 55]. BotNet [43] relates the transformer

block with the fully attentional version of bottleneck block in ResNet [19].

2.3. Combining Self-Attention and Convolution

There are four categories of methods to combine the self-attention and convolution. *Approaches in different categories can be used together.*

The first one is using depth-wise convolution to replace the position embedding layer in self-attention, including CVPT [12] and CoAT [57]. The second one is the serial connection with convolution before attention. CvT [54] applies convolutions before calculating the query, key and value in self-attention. The third one is the serial connection with attention before convolution. Convolutions are applied after the self-attention layer, including LocalViT [28] and PVTv2 [52]. The final one is parallel connection. Attention Augmentation [2] augments the convolution features with attention features. It is intended to incorporate the long-range connections into the convolutions. Our approach, LESA falls into this category. Different from Attention Augmentation, it enhances the local term in self-attention.

3. LESA: Locally Enhanced Self Attention

3.1. Decomposition of Self-Attention

We decompose the self-attention into local and context terms. Let $x \in \mathbb{R}^{d_{in} \times H \times W}$ be the input, where d_{in} is the feature channels and H, W are the height and width in spatial dimensions. In this case, each pixel is connected with all the other pixels in the computation. We consider the all-to-all self-attention since it has been adopted as a building layer and shows superior performance [43, 51]. Specifically, we can write the formula of self-attention as:

$$x_{i,j}^{l+1} = \sum_{(h,w)=(1,1)}^{(H,W)} \text{softmax}(q_{i,j}^\top k_{h,w} + q_{i,j}^\top r_{(i,j) \rightarrow (h,w)}) v_{h,w} \quad (1)$$

where i, j and h, w represent the spatial locations of the pixel and l specifies the layer index. $q, k, v = xW_q, xW_k, xW_v$ are the query, key and value which are obtained by applying three different 1×1 convolutions on x^l . $W_q, W_k \in \mathbb{R}^{d_{in} \times d_{qk}}$ and $W_v \in \mathbb{R}^{d_{in} \times d_{out}}$ are learnable parameters, where d_{qk}, d_{out} are intermediate and output channels. $r_{(i,j) \rightarrow (h,w)}$ is the relative position embedding, and for simplicity we will use the notation $r_{h,w}$. This formula shows the activation $x_{i,j}^{l+1}$ integrates the information conveyed by all the pixels $x_{i,j}^l$. To comprehend this operation, we decompose the information flow and reformulate the equation as the combination of local term and context

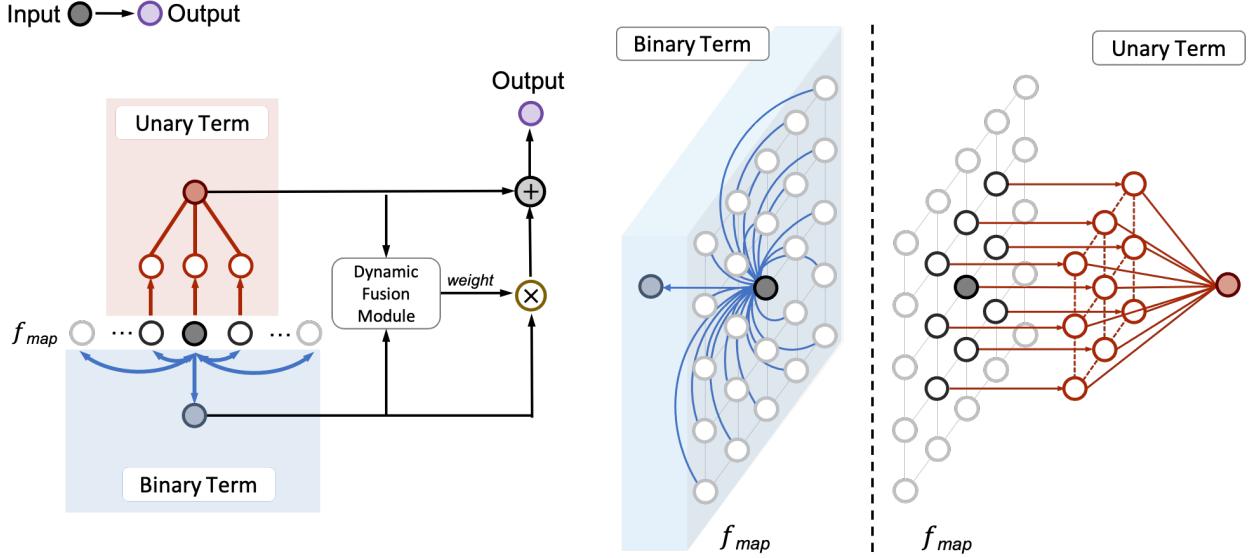


Figure 3. Visualizing Locally Enhanced Self-Attention (LESA) at one spatial location. The left part is the operation pipeline. f_{map} is the feature map. The red and blue arrows represent the unary and binary operations.

term:

$$x_{i,j}^{l+1} = S_{H,W}^{i,j} (q_{i,j}^\top k_{i,j} + q_{i,j}^\top r_{i,j}) v_{i,j} + \sum_{(h,w) \neq (i,j)}^{(H,W)} S_{H,W}^{h,w} (q_{i,j}^\top k_{h,w} + q_{i,j}^\top r_{h,w}) v_{h,w} \quad (2)$$

$$S_{H,W}^{h,w}(z) = \frac{e^{z_{h,w}}}{\sum_{(m,n)}^{(H,W)} e^{z_{m,n}}} \quad (3)$$

For the spatial location (i, j) , the first local term computes activation by looking at itself, while the second context term computes activation by looking at others. The softmax generates the weights of contribution. Through this decomposition, we can interpret self-attention as a double-source feature extractor, which consists of a pair of unary and binary terms. Unary and binary terms are computed by the queries, keys, and values q, k, v at different spatial locations with *shared* projection matrices W_q, W_k, W_v . Consequently, the outputs *entangle* the local and context features.

We perform an ablation study to investigate the contribution of these two terms. Specifically, we take a ResNet50 [19] and replace the convolution layers of its last two stages with self-attention. The model is trained from scratch on ImageNet [39]. During inference, we track the softmax operations of all self-attention layers and obtain the weights for the unary and binary terms, $S_{H,W}^{h,w}$ and $\sum_{(h,w) \neq (i,j)}^{(H,W)} S_{H,W}^{h,w}$ whose summation equals to 1. By averaging them across all layers, we obtain the weighted contributions of these two terms. Then, we ablate the unary in the evaluation phase. The results are shown in Tab. 1.

Methods	Top-1 Err. (%)	Weight Pct. (%)
SA	21.31	100.00
SA - unary term	28.87	98.03

Table 1. Contributions of the unary term in Self-Attention (SA). We replace the spatial convolutions in the 3rd and 4th stages of ResNet50 [19] with self-attention. By tracking the softmax operations, we record the weights of the unary and binary terms $S_{H,W}^{h,w}$ and $\sum_{(h,w) \neq (i,j)}^{(H,W)} S_{H,W}^{h,w}$ in Eqn. (2). They add up to 1 at each layer. The weight percentage is the average across all the layers. We observe that the unary term is important. The removal of unary terms whose weight percentage is less than 2% increases the error rate by 7.56% (or > 35% relative increase).

We can observe that self-attention is mainly contributed by the binary operations, but the unary term is also important. Although the weights of unary terms only take less than 2%, the removal of which causes 7.56% drop of accuracy or 35% relative increase on the error rate. When analyzing the self-attention by this decomposition, unary term plays a significant role, but most of the computations and focuses are given to binary operations.

3.2. Locally Enhanced Self-Attention

Local and context terms have been long used in formulating the graphical models for vision tasks, such as image denoising, segmentation, and surface reconstruction [36]. The fully connected Conditional Random Fields (CRFs) have been introduced on top of the deep networks for semantic segmentation [8]. It aims at coupling the recognition

capacity and localization accuracy, and achieves excellent performance. For a grid of pixels in the form of a graph $G = (V, E)$, the energy to be minimized for the CRF is defined by:

$$\phi_c = \sum_{x \in V} \Phi_U(x) + \sum_{(x_u, x_v) \in E} \Phi_B(x_u, x_v) \quad (4)$$

where u, v indicate the different vertices in V . The unary term is $\Phi_U(x) = -\log(P(x))$ where $P(x)$ is the probability of assigning x the ground truth label by the model. The binary term is $\Phi_B(x) = \pi((x_u, p_u), (x_v, p_v)) + \pi(p_u, p_v)$ where x and p are the contents and spatial positions. π is the probability density function to measure the similarity of two values, which can be chosen as Gaussian.

The unary term is utilized for recognition while the binary term for spatial and content interactions. Inspired by these and our decomposition analysis, we propose Locally Enhanced Self-Attention (LESA). It contains a unary term incorporated with convolutions, and a binary term for feature interactions. Locally Enhanced Self-Attention is defined by

$$x_{i,j}^{l+1} = m_{i,j} + \omega_{i,j} \sum_{(h,w)=(1,1)}^{(H,W)} S_{H,W}^{h,w} q_{i,j}^\top k_{h,w} v_{h,w} \quad (5)$$

$$m = x W_g^k W_1^1 \quad (6)$$

where $\omega_{i,j}$ is the weight that is shown in Equ. (7). m is the local term obtained by applying two consecutive convolutions. $W_g^k \in \mathbb{R}^{d_{in} \times d_{out}}$ is a learnable matrix where k and g represent the spatial extent and group number of the convolution. $W_1^1 \in \mathbb{R}^{d_{out} \times d_{out}}$ is a learnable projection matrix representing 1×1 convolution. By this design, the multi-head mechanism is integrated. $m_{i,j}$ is the unary activation at spatial location (i, j) . This formulation of LESA also enables us to change W_g^k to deformable convolution [14] for the tasks of object detection and instance segmentation. Self-attention focuses on the binary operations. We use it as the context term to model the feature interactions with relative spatial relationships among all possible pairs of pixels.

3.3. Dynamic Fusion of the Unary and Binary Terms

Adding the unary and binary terms is a static way of merging the two terms with equal weights. A more flexible strategy is to allocate the weights on demand under different circumstances. For example, in object detection, the locality of pixel dependencies is more important than the context when detecting multiple small objects in an image. We achieve a dynamic control by multiplying the binary term by ω and adaptively adjusting the relative weights of the two terms, which is shown in Equ. (5). Specifically, we can

Models	Ops.	Pms (M)	Acc. (%)		Weights (%)	
			T-1	T-5	Unary	Binary
R50	Conv.	25.6	76.1	92.9	100	0.0
	SA	18.1	78.7	94.2	2.0	98.0
	LESA	23.8	79.6	94.8	67.0	33.0
WR50	Conv.	68.9	78.5	94.1	100	0.0
	SA	37.7	79.7	94.7	2.5	97.5
	LESA	60.6	80.2	95.1	66.9	33.2

Table 2. ImageNet classification results. We replace the spatial convolutions in the 3rd and 4th stages of ResNet [19] and WRN [58] with self-attention and LESA. We employ the position embedding method used in [51]. For self-attention, the weight percentages are obtained by tracking softmax operations which is the same as Tab. 1. For LESA, the ω in Equ. (7), the output of the dynamic fusion module is tracked. $\frac{1}{1+\omega}$ and $\frac{\omega}{1+\omega}$ are the weights for the unary and binary terms in LESA, which are then averaged across the spatial locations, feature channels, and layers as the final weight percentage. We can observe that compared with self-attention, LESA achieves a more balanced utilization of both the unary and binary terms. In both the top-1 and top-5 accuracy, LESA outperforms other baselines. *LESA has a more significant improvement in object detection where the local cues are particularly important.*

write the formula of $\omega_{i,j}$ as:

$$\omega_{i,j} = \text{sigmoid}\left(F(m_{i,j} \parallel \sum_{(h,w)=(1,1)}^{(H,W)} S_{H,W}^{h,w} q_{i,j}^\top k_{h,w} v_{h,w})\right) \quad (7)$$

where $\omega \in \mathbb{R}^{d_{out} \times H \times W}$ and $\omega_{i,j}$ corresponds to one spatial location. $F : \mathbb{R}^{2d_{out} \times H \times W} \mapsto \mathbb{R}^{d_{out} \times H \times W}$ is a function. Sigmoid operation is performed element-wisely on the logits given by F , making ω range from 0 to 1. Regarding F , we design it as a three-layer perceptron and adopt the pre-activation design [20]. Concretely, together with sigmoid we can represent the pipeline as $F + \text{sigmoid} : \text{BN} \rightarrow \text{relu} \rightarrow fc_1 \rightarrow \text{BN} \rightarrow \text{relu} \rightarrow fc_2 \rightarrow \text{BN} \rightarrow \text{sigmoid}$ where BN is batch normalization layer [26], and $fc_1 : \mathbb{R}^{2d_{out}} \mapsto \mathbb{R}^{d_{out}}$, $fc_2 : \mathbb{R}^{d_{out}} \mapsto \mathbb{R}^{d_{out}}$ are two fully connected layers. The position embedding is omitted in Equ. 5 and 7 for simplicity. In our design, ω depends on the contents of the unary and binary terms and controls their relative weights at different spatial locations and in different feature channels.

4. Experiments

4.1. ImageNet Classification

- **Settings** We perform image classification experiments on ILSVRC2012 [39] which is a popular subset of the ImageNet database [15]. There are 1.3M images in the training

BB.	Ops.	Epo.	Object Detection						Intancecs Segmentation					
			AP ^b	AP ₅₀	AP ₇₅	AP _S	AP _M	AP _L	AP ^m	AP ₅₀	AP ₇₅	AP _S	AP _M	AP _L
R50	Conv.	12	37.8	58.6	41.0	22.1	41.1	48.3	34.5	55.5	36.7	18.6	37.5	45.8
R50	SA	12	38.5	60.0	41.6	23.7	41.4	49.5	34.7	56.7	36.6	19.7	37.5	46.4
R50	LESA	12	40.3	61.2	43.6	24.7	43.3	52.0	36.5	58.3	38.9	21.0	39.5	48.9
WR50	Conv.	12	40.9	61.8	44.8	24.2	44.7	52.5	36.9	58.6	39.6	20.3	40.5	49.0
WR50	SA	12	42.0	63.2	45.6	25.4	45.0	54.6	37.6	60.1	40.1	21.1	40.5	50.4
WR50	LESA	12	44.8	66.1	49.1	27.8	48.3	58.4	39.9	62.9	42.8	22.7	43.1	54.0
R50	Dconv.	20	43.0	64.0	47.3	25.8	46.4	56.4	38.6	61.0	41.2	21.4	41.8	52.4
R50	LESA	20	44.0	64.8	48.0	26.3	47.9	57.3	39.2	61.8	41.9	21.9	42.8	52.5
WR50	Dconv.	20	44.7	65.5	49.1	27.0	48.2	59.3	39.8	62.5	42.7	22.6	43.1	54.7
WR50	LESA	20	46.3	67.5	50.7	28.8	49.9	60.2	40.9	64.3	43.9	23.5	44.5	54.9
WR50	HTC Dconv.	20	47.4	66.7	51.7	30.0	50.9	61.8	41.9	64.0	45.4	24.2	45.3	57.1
WR50	LESA	20	48.8	68.4	53.3	31.7	52.3	63.1	43.0	65.7	46.4	25.8	46.4	57.8
WR50	HTC Dconv ^b	20	48.5	67.8	52.9	31.3	52.0	62.3	42.8	65.6	46.5	25.6	46.1	57.1
WR50	LESA_H	20	50.1	69.7	54.4	32.3	53.6	65.0	43.9	67.1	47.5	26.6	47.1	59.1

Table 3. COCO object detection and instance segmentation on val2017. We use Mask-RCNN [18] and HTC [5] frameworks and employ FPN [29] on the backbones ResNet [19] and WRN [58]. Dconv. stands for deformable convolutions. We adopt two standard training schedules that have 12 and 20 epochs. The learning rate is adjusted 10 times smaller after 8, 11 epochs and 16, 19 epochs, respectively. The images are resized to 1024×1024 by default, and the superscript ^b (higher resolution) indicates that they are resized to 1280×1280 . We can observe that LESA outperforms the convolution, self-attention, and deformable convolution baselines in all the experiments.

BB.	Ops.	Epo.	Object Detection						Intancecs Segmentation					
			AP ^b	AP ₅₀	AP ₇₅	AP _S	AP _M	AP _L	AP ^m	AP ₅₀	AP ₇₅	AP _S	AP _M	AP _L
R50	LESA	12	40.5	61.7	44.3	24.2	42.4	50.1	36.6	58.7	39.0	20.6	38.3	46.9
WR50	LESA	12	44.8	66.4	49.2	26.9	47.8	56.0	40.0	63.1	43.1	22.7	42.6	51.8
R50	LESA	20	44.2	65.5	48.3	26.4	46.9	55.5	39.6	62.4	42.6	22.5	42.0	51.1
WR50	LESA	20	46.5	67.7	51.0	28.4	49.4	58.3	41.2	64.6	44.5	23.9	43.7	53.4
WR50	HTC LESA	20	49.0	68.8	53.4	30.3	51.8	61.7	43.4	66.4	47.1	25.2	45.9	56.6
WR50	HTC LESA ^b	20	50.5	70.2	54.9	31.8	53.2	63.3	44.4	67.7	48.2	26.6	46.7	57.5

Table 4. COCO object detection and instance segmentation on test-dev2017 for the models in Tab. 3.

set and 50K images in the validation set. In total, it includes 1,000 object classes.

ResNet [19], a family of canonical models or backbones for vision tasks, and its larger variant WRN [58] are used to study LESA. There are 4 stages in ResNet and each one is formed by a series of bottleneck blocks. ResNet50 can be represented by the bottleneck numbers $\{3, 4, 6, 3\}$. We replace the $\text{conv}3 \times 3$ in the bottleneck with the self-attention and LESA. The kernel channels of these $\text{conv}3 \times 3$ in WRN are twice as large as those in ResNet.

We perform the replacement in the last two stages, which is enough to show the advantages of LESA. For convolution baselines, we use Torchvision official models [34]. For self-attention baselines and LESA, we set head number 8 for both of them and train the models from scratch. We

set the stride of the last stage to be 1 following [43]. We keep the first bottleneck in stage 3 unchanged which has the stride 2 convolution. We employ a canonical training scheme with 5 linear warm-up and 90 training epochs with a batch size 128. Following [38, 51], we employ SGD with Nesterov momentum [32, 44] and cosine annealing learning rate initialized as 0.05. The experiments are performed on 4 NVIDIA TITAN XP graphics cards.

- **Results** The results are summarized in Tab. 2. For both the top-1 and top-5 accuracy, LESA surpasses the convolution and self-attention baselines. Our dynamic fusion module controls the binary term using ω in Equ. (7), and thus the weights for the unary and binary terms are $\frac{1}{1+\omega}$ and $\frac{\omega}{1+\omega}$, respectively. As ω is dependent on the inputs, spatial locations, and feature channels, we average the weights

across them in our records. In self-attention, the weights are calculated by softmax operations as used in Tab. 1. It is observed that the weight distribution in self-attention are imbalanced. The unary term only has a weight percentage less than 3%, more than 32 times smaller than the binary term’s. While for LESA, their weight percentages are about 67% and 33%, respectively. In the tasks of object detection where local cues are particularly important, LESA shows better improvement, which is shown in Tab. 3 and 4.

4.2. COCO Object Detection and Instance Segmentation.

- **Settings** We perform object detection experiments on COCO dataset [30] and use the 2017 dataset splits. There are 118K images in the training set and 5K images in the validation set. In total there are 80 object categories and on average each image contains 3.5 categories and 7.7 instances.

The widely used Mask-RCNN [18] and HTC [5] with the backbones equipped with FPN [29] are used to study LESA for object detection and instance segmentation. We use mmdetection [6] as the codebase. The ImageNet pre-trained checkpoints are utilized to initialize the backbones. There are 5 stages in the ResNet-FPN and the output strides are $\{4, 8, 16, 32, 64\}$. We replace the spatial convolutions in the 3rd and 4th stages. The images are resized to 1024×1024 and 1280×1280 in the experiments. Since the image size in classification is 224×224 , we initialize new position embedding layers used in [38, 41]. For training, we employ the **1×** and **20e** schedules. The total epochs and epochs after which the learning rate is multiplied by 0.1 are $\{12, 8, 11\}$ and $\{24, 16, 19\}$, respectively. For the HTC framework, we employ multi-scale training for both the baseline and our method: with 0.5 probability that both sides of the image are resized to a scale uniformly chosen from 512 to 1024, and with another 0.5 probability to a scale that is uniformly sampled from 1024 to 2048. Mask-RCNN does not use multi-scale training.

We also study adopting the deformable unary terms in LESA. Specifically, we replace W_g^k in Equ. (6) to deformable convolutions [14]. We set the group number of offsets as 1. Following the standard setting [37], the convolutions in the 2nd stage in both the baselines and our models are also replaced with deformable convolutions. Our experiments with Mask-RCNN framework are performed on 4 NVIDIA TITAN XP graphics cards and those with HTC framework on 4 TITAN RTX graphics cards.

- **Results** The results are summarized in Tab. 3 and 4. We use the same testing pipeline for `val2017` and `test-dev2017`. LESA provides the best bounding box mAP and mask mAP for all the small, medium, and large objects compared with convolution, self-attention, and deformable convolution baselines in all scenarios.

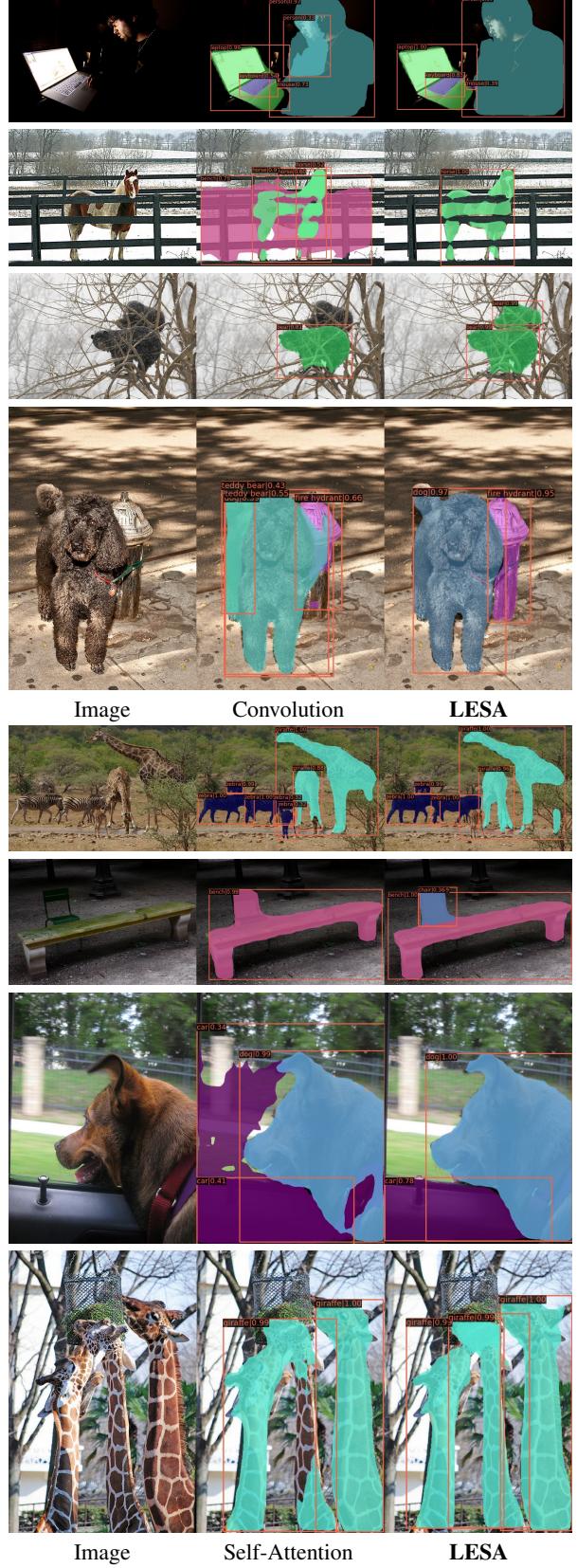


Figure 4. COCO object detection and instance segmentation.

Categories	Methods
Conv. as position embedding of SA	CVPT; CoaT
Serial Conn. with Conv. before SA	CvT;
Serial Conn. with Conv. after SA	LocalViT; PVTv2
Parallel Conn.	AA; static LESA; LESA

Table 5. Categories of merging methods for convolution and self-attention(discussed in the related work section). Conn. stands for connection. Approaches in different categories can be used together. As a parallel connection method, LESA is compared with AA (Attention Augmentation) in Tab. 6, and shows the superior performances.

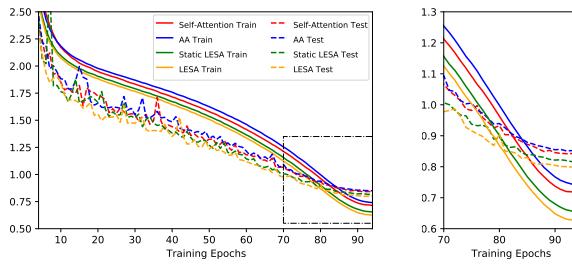


Figure 5. Training and testing curves of ablation studies. AA stands for Attention Augmentation [2]. The right figure shows the details in the rectangular region of the left figure.

5. Relationships with other methods

- Relationship with other enhanced self-attention methods** LESA is proposed as a combination mechanism for convolution and self-attention, which can be used with other enhanced self-attention together.
- Relationships with other merging methods for self-attention and convolution** As summarized in Tab. 5 and discussed in related work section, there are four categories, and approaches in different ones can be used together. LESA falls into the category of parallel connection. We compare LESA with Attention Augmentation [2] and meanwhile, perform the ablation studies. Static LESA stands for adding the unary and binary terms of LESA without the dynamic fusion module.

The experiments are performed on ImageNet. The setting is the same as the main experiments. The results are summarized in Tab. 6. It is observed the performance difference between Attention Augmentation and self-attention is marginal. Both Static LESA and LESA show better performances than Attention Augmentation with less parameters. LESA has stronger performance than the static one, demonstrating effectiveness of the dynamic fusion module.

Operations	Params (M)	Accuracy (%)	
		Top-1	Top-5
Convolution	25.56	76.13	92.86
Self-Attention	18.06	78.69	94.22
AA [2]	25.15	78.74	94.23
Static LESA	20.43	79.12	94.62
LESA	23.78	79.55	94.79

Table 6. Parallel connection methods and ablation studies. The experiments are performed with ResNet50 [19] on ImageNet. The spatial convolution layers in last two stages are replaced. AA stands for Attention Augmentation [2], which is another parallel connection method of combining convolution and attention. Static LESA stands for adding the unary and binary terms without the dynamic fusion module. Both Static LESA and LESA achieve better performances than AA with less parameters. The effectiveness of the dynamic fusion module can be proved by the superiority of LESA over Static LESA.

6. Conclusion

From the perspective of fully connected Conditional Random Fields (CRFs), we decouple the self-attention into the local and context terms. They are the unary and binary terms that are calculated by the queries, keys and values in the attention mechanism. However, there lacks distinction between the local and context cues as they are obtained by using the same set of projection matrices. In addition, we observe the contribution of the local terms is very small which is controlled by the softmax operation. By contrast, the standard Convolutional Neural Networks (CNNs) show excellent performances in various vision tasks and rely solely on the local terms.

In this work, we propose Locally Enhanced Self-Attention (LESA). First, we enhance the unary term by incorporating it with convolutions. The multi-head mechanism is realized by using grouped convolution followed by the projection layer. Second, we propose a dynamic fusion module to combine the unary and binary terms with input-dependent relative weights. We demonstrate the superiority of LESA over the convolution and self-attention baselines in the tasks of ImageNet classification, and COCO object detection and instance segmentation.

Limitations LESA shares a common limitation with self-attention, which is the large memory consumption. This is due to the large dimensions of the similarity matrix which is computed by the queries and keys and where the softmax operation is applied. Our future works include designing a LESA that consumes small memory but still has the great power of capturing the context information. This will also address the common memory issue in other self-attention models.

References

- [1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014. 3
- [2] Irwan Bello, Barret Zoph, Ashish Vaswani, Jonathon Shlens, and Quoc V Le. Attention augmented convolutional networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3286–3295, 2019. 3, 8
- [3] Antoni Buades, Bartomeu Coll, and J-M Morel. A non-local algorithm for image denoising. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, pages 60–65. IEEE, 2005. 3
- [4] 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, 3
- [5] Kai Chen, Jiangmiao Pang, Jiaqi Wang, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jianping Shi, Wanli Ouyang, et al. Hybrid task cascade for instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4974–4983, 2019. 6, 7
- [6] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, et al. Mmdetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019. 7
- [7] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062*, 2014. 3
- [8] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017. 4
- [9] Liang-Chieh Chen, Huiyu Wang, and Siyuan Qiao. Scaling wide residual networks for panoptic segmentation. *arXiv preprint arXiv:2011.11675*, 2020. 3
- [10] Yunpeng Chen, Yannis Kalantidis, Jianshu Li, Shuicheng Yan, and Jiashi Feng. a^2 -nets: Double attention networks. *arXiv preprint arXiv:1810.11579*, 2018. 3
- [11] François Fleuret. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017. 3
- [12] Xiangxiang Chu, Zhi Tian, Bo Zhang, Xinlong Wang, Xiaolin Wei, Huaxia Xia, and Chunhua Shen. Conditional positional encodings for vision transformers. *arXiv preprint arXiv:2102.10882*, 2021. 3
- [13] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014. 1
- [14] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 764–773, 2017. 3, 5, 7
- [15] J. Deng, W. Dong, R. Socher, L.J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition*, 2009. 5
- [16] Yihe Dong, Jean-Baptiste Cordonnier, and Andreas Loukas. Attention is not all you need: Pure attention loses rank doubly exponentially with depth. *arXiv preprint arXiv:2103.03404*, 2021. 2
- [17] 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, 3
- [18] 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. 2, 6, 7
- [19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2, 3, 4, 5, 6, 8
- [20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European conference on computer vision*, pages 630–645. Springer, 2016. 5
- [21] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 1
- [22] 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
- [23] Han Hu, Zheng Zhang, Zhenda Xie, and Stephen Lin. Local relation networks for image recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3464–3473, 2019. 3
- [24] 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. 3
- [25] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 3
- [26] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR, 2015. 5
- [27] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012. 3
- [28] Yawei Li, Kai Zhang, Jiezhang Cao, Radu Timofte, and Luc Van Gool. Localvit: Bringing locality to vision transformers. *arXiv preprint arXiv:2104.05707*, 2021. 3

- [29] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. [2](#), [6](#), [7](#)
- [30] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. [2](#), [7](#)
- [31] 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. [1](#), [3](#)
- [32] Yurii E Nesterov. A method for solving the convex programming problem with convergence rate $O(1/k^2)$. In *Dokl. akad. nauk Sssr*, volume 269, pages 543–547, 1983. [6](#)
- [33] George Papandreou, Iasonas Kokkinos, and Pierre-André Savalle. Modeling local and global deformations in deep learning: Epitomic convolution, multiple instance learning, and sliding window detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 390–399, 2015. [3](#)
- [34] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703*, 2019. [6](#)
- [35] Florent Perronnin, Jorge Sánchez, and Thomas Mensink. Improving the fisher kernel for large-scale image classification. In *European conference on computer vision*, pages 143–156. Springer, 2010. [3](#)
- [36] Simon JD Prince. *Computer vision: models, learning, and inference*. Cambridge University Press, 2012. [4](#)
- [37] Siyuan Qiao, Liang-Chieh Chen, and Alan Yuille. Detectors: Detecting objects with recursive feature pyramid and switchable atrous convolution. *arXiv preprint arXiv:2006.02334*, 2020. [3](#), [7](#)
- [38] Prajit Ramachandran, Niki Parmar, Ashish Vaswani, Irwan Bello, Anselm Levskaya, and Jonathon Shlens. Stand-alone self-attention in vision models. *arXiv preprint arXiv:1906.05909*, 2019. [1](#), [3](#), [6](#), [7](#)
- [39] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015. [2](#), [4](#), [5](#)
- [40] Jorge Sánchez and Florent Perronnin. High-dimensional signature compression for large-scale image classification. In *CVPR 2011*, pages 1665–1672. IEEE, 2011. [3](#)
- [41] Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. *arXiv preprint arXiv:1803.02155*, 2018. [7](#)
- [42] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. [3](#)
- [43] 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. [1](#), [3](#), [6](#)
- [44] Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *International conference on machine learning*, pages 1139–1147. PMLR, 2013. [6](#)
- [45] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015. [3](#)
- [46] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114. PMLR, 2019. [3](#)
- [47] 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. [3](#)
- [48] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017. [1](#), [3](#)
- [49] Ziyu Wan, Jingbo Zhang, Dongdong Chen, and Jing Liao. High-fidelity pluralistic image completion with transformers. *arXiv preprint arXiv:2103.14031*, 2021. [1](#)
- [50] Huiyu Wang, Yukun Zhu, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Max-deeplab: End-to-end panoptic segmentation with mask transformers. *arXiv preprint arXiv:2012.00759*, 2020. [1](#)
- [51] Huiyu Wang, Yukun Zhu, Bradley Green, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Axial-deeplab: Stand-alone axial-attention for panoptic segmentation. In *European Conference on Computer Vision*, pages 108–126. Springer, 2020. [1](#), [3](#), [5](#), [6](#)
- [52] Wenhui Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pvttv2: Improved baselines with pyramid vision transformer. *arXiv preprint arXiv:2106.13797*, 2021. [3](#)
- [53] 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](#)
- [54] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. Cvlt: Introducing convolutions to vision transformers. *arXiv preprint arXiv:2103.15808*, 2021. [3](#)
- [55] Zhanghao Wu, Zhijian Liu, Ji Lin, Yujun Lin, and Song Han. Lite transformer with long-short range attention. *arXiv preprint arXiv:2004.11886*, 2020. [3](#)
- [56] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017. [3](#)

- [57] Weijian Xu, Yifan Xu, Tyler Chang, and Zhuowen Tu. Co-scale conv-attentional image transformers. *arXiv preprint arXiv:2104.06399*, 2021. [3](#)
- [58] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016. [2](#), [3](#), [5](#), [6](#)
- [59] Hengshuang Zhao, Jiaya Jia, and Vladlen Koltun. Exploring self-attention for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10076–10085, 2020. [3](#)
- [60] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets v2: More deformable, better results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9308–9316, 2019. [3](#)