

Spatial Bias for Attention-free Non-local Neural Networks

Junhyung Go^b, Jongbin Ryu^{a,b}

^a*Department of Computer Engineering, Ajou University*

^b*Department of Artificial Intelligence, Ajou University*

Abstract

In this paper, we introduce the spatial bias to learn global knowledge without self-attention in convolutional neural networks. Owing to the limited receptive field, conventional convolutional neural networks suffer from learning long-range dependencies. Non-local neural networks have struggled to learn global knowledge, but unavoidably have too heavy a network design due to the self-attention operation. Therefore, we propose a fast and lightweight spatial bias that efficiently encodes global knowledge without self-attention on convolutional neural networks. Spatial bias is stacked on the feature map and convolved together to adjust the spatial structure of the convolutional features. Therefore, we learn the global knowledge on the convolution layer directly with very few additional resources. Our method is very fast and lightweight due to the attention-free non-local method while improving the performance of neural networks considerably. Compared to non-local neural networks, the spatial bias use about $\times 10$ times fewer parameters while achieving comparable performance with 1.6 ~ 3.3 times more throughput on a very little budget. Furthermore, the spatial bias can be used with conventional non-local neural networks to further improve the performance of the backbone model. We show that the spatial bias achieves competitive performance that improves the classification accuracy by +0.79% and +1.5% on ImageNet-1K and cifar100 datasets. Additionally, we validate our method on the MS-COCO and ADE20K datasets for downstream tasks involving object detection and semantic segmentation.

Keywords: Non-local operation, Long-range dependency, Spatial bias, Global context, Image classification, Convolutional neural networks

1. Introduction

Convolutional neural networks (CNNs) excel in extracting nuanced local information. Thanks to this advantage, CNNs are utilized for a variety of visual recognition tasks. However, their inability to effectively capture the global context has been mentioned numerous repeatedly. Due to the limited receptive field size, the convolution focuses on a small region that learns only local information; to overcome this, several approaches to increase the receptive field size have been extensively studied, such as building deeper layers He et al. (2016), employing different kernel sizes Szegedy et al. (2015); Li et al. (2019); Li and Zhang (2022), and learning non-local pixel-level pairwise relationships Wang et al. (2018); Cao et al. (2019); Fang et al. (2021); You et al. (2022); Ding et al. (2023); Cho et al. (2022); Chi et al. (2020); Xie et al. (2022). Among these methods, self-attention based non-local neural networks Wang et al. (2018) have been a major approach to capture long-range information. However, they exploit an excessive amount of resources because of the self-attention operation. Therefore, this paper presents a novel lightweight method that directly learns the long-range dependency during the convolution operation. The proposed method acquires global knowledge by

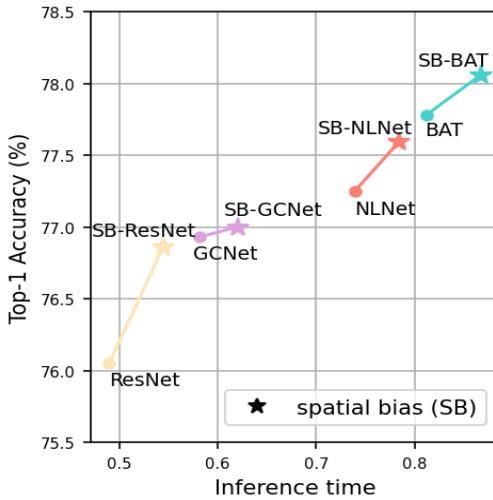


Figure 1: Performance comparison for the proposed method and conventional non-local neural networks on ImageNet-1K dataset. • denotes the naive ResNet backbone and conventional non-local neural networks and ★ present performance improvement of the conventional networks with our spatial bias. In all cases, the proposed spatial bias enhance networks with minimal computational complexity.

incorporating a spatial bias term into the convolution operation. A spatial bias with long-range correlation is added to the position in which convolution is performed.

The proposed method allows for the simultaneous learning of local information from the convolution term and global knowledge from the spatial bias term. In addition, a minimal amount of resources are used for the proposed spatial bias term, and thus our method is very fast and lightweight compared to the conventional self-attention based non-local neural networks. We extensively carry out experiments on the number of parameters, FLOPs, throughput, and accuracy to show the efficacy of the proposed spatial bias method. As shown in Fig. 1, the inference time overhead of our spatial bias is **1.6** to **3.3** times less than that of non-local neural network while achieving competitive performance compared to the non-local networks Wang et al. (2018); Cao et al. (2019); Chi et al. (2020). The proposed spatial bias further improves the performance of backbone networks in conjunction with existing self-attention operations. The following is a summary of the contributions regarding our spatial bias.

- We introduce a spatial bias that takes into account both local and global knowledge in a convolution operation. Thanks to the proposed lightweight architecture, the spatial bias term is computed very quickly and with a small amount of overhead.
- We show that the proposed spatial bias term significantly improves the performance while incurring very modest overheads: in the case of ResNet-50 backbone, the parameter overhead of spatial bias has only 6.4% and $\times 3.3$ faster compared to the non-local neural network (NLNet) Wang et al. (2018).
- We verify the generalizability of the proposed spatial bias by combining it with other non-local methods and backbones. We also confirm that spatial bias improves performance in downstream tasks.

2. Related Work

Non-local Neural Network with Self-attention. The non-local neural networks Wang et al. (2018); Cao et al. (2019); Chi et al. (2020) using self-attention operation that learns long-range dependency has been most widely studied. Unlike

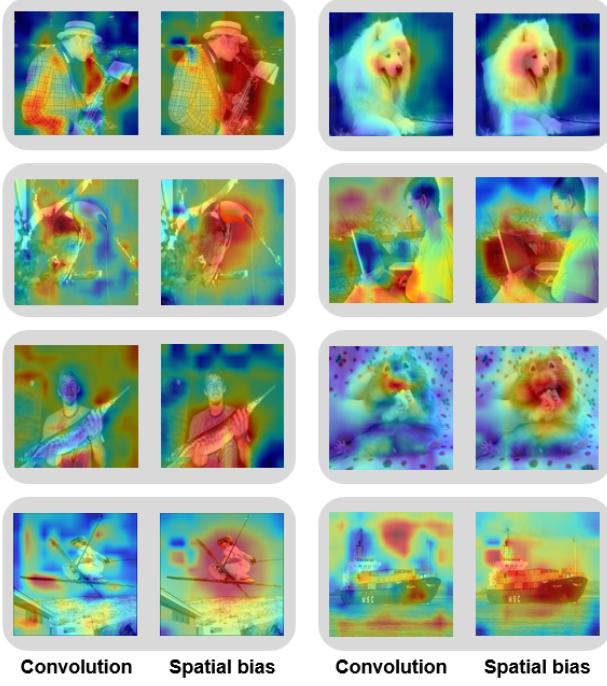


Figure 2: Grad-Cam Selvaraju et al. (2017) visualization for spatial bias and convolution feature map. Notably, the grad-cam on the spatial bias exposes wider regions, which aids in learning global knowledge.

convolution operation, self-attention learns global knowledge in a single layer, which alleviates the narrow receptive field problem of CNNs. This approach performs well when applied to a variety of visual tasks. In particular, NLNet Wang et al. (2018) is the first study to exploit the self-attention operation for learning the pairwise relationship at the global pixel-level. However, since NLNet is calculated after obtaining an attention map independent of each query, the computation complexity is very high. For different query locations, GCNet Cao et al. (2019) generates similar attention maps, thereby modeling an efficient non-local block. In order to create lighter non-local networks, CC-Net Huang et al. (2019), A2Net Chen et al. (2018), and BAT Chi et al. (2020) have been introduced by using an iterative cross-attention module Huang et al. (2019), dual-attention method Chen et al. (2018), and data-adaptive bilinear attentional transformation Chi et al. (2020). Fang *et al.* Fang et al. (2021) proposes an location-based attention method that distills positive information in a global context. Recently, the non-local neural networks are

used to various tasks such as histopathological image classification Ding et al. (2023) and hand detection Xie et al. (2022). These methods have contributed to the design paradigm of non-local neural networks with the reduced overhead of self-attention operation. However, we argue that they still suffer from a fatal flaw in that their computational cost is quadratic $O(n^2)$ ¹ which causes a slowdown of inference time.

Due to the heavy design of self-attention, conventional non-local methods are inserted only at specific layers in a convolutional neural network by consideration of the throughput and network size. Additionally, since traditional non-local operations only consider spatial correlation by merging channels, they are blind to any channel correlation. Therefore, to overcome these limitations, we propose spatial bias, an attention-free non-local neural network with a fast and lightweight architecture. In comparison to self-attention based non-local neural networks, the proposed spatial bias achieves comparable performance with $1.6 \sim 3.3$ times more throughput on a very little budget. Additionally, lightweight spatial bias can be employed across all network levels, complementing the existing heavy self-attention that can be given to particular layers. Thus, our spatial bias enables a network to learn more about global knowledge due to its fast and lightweight properties.

Architecture Modeling. Recently, effective neural networks have shown advances across a range of many visual recognition tasks. The modern CNN architecture conceived by LeNet LeCun et al. (1998) was realized a decade ago by AlexNet Krizhevsky et al. (2012), and various studies have been conducted to improve its performance and usefulness. Since then, CNN Simonyan and Zisserman (2014) with small filter sizes has been developed to encode more precise regional data. Consequently, skip connections have made it possible to build deeper networks He et al. (2016) and several studies have been done to increase expressiveness by varying the width, multi-path block design, or grouped feature flow of neural networks Zagoruyko and Komodakis (2016); Szegedy et al. (2015); Xie et al. (2017); Gao and Zhou (2023); Schwarz Schuler et al. (2022). Through a multi-branch design, recent CNN architectures Szegedy et al. (2015); Li et al. (2019); Zhang et al. (2020); Gao et al. (2019); Guo et al. (2021) communicate information between branches. Additionally, several methods Wang et al. (2018); Chi et al. (2020); Cao et al. (2019) capture long-range dependencies by taking advan-

¹ n indicates the number of all positions in the feature map

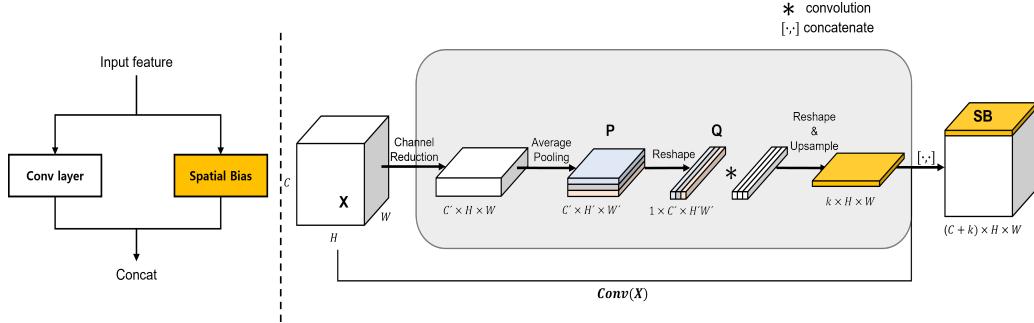


Figure 3: Design of Spatial Bias term. The figure on the left shows the overall workflow and the right one shows its detail. In the right figure, to capture the global dependency, the channel and spatial size of the feature map are reduced through the 1×1 convolution and average pooling operations. We aggregate spatial bias on a reduced feature map using a simple 1D convolution operation.

tage of the self-attention operation that guarantees a better understanding of global knowledge for the visual recognition task.

3. Proposed Method

The convolution utilizes a shared weight within a limited local window, allowing CNN to have the property of translational equivalence Zhang (2019). Recently, this property has been identified as the inductive bias Baxter (2000), and it has been stated repeatedly that convolution is not particularly effective at learning the relationship between long-range objects Wang et al. (2018). To address this problem, we propose a spatial bias term to compensate for these shortcomings in the convolution. Different from the existing method Wang et al. (2018); Cao et al. (2019); Chi et al. (2020) using the heavy self-attention module, the proposed method learns the global knowledge by adding a few spatial bias channels to the convolutional feature map. Inspired by parallel network designs Szegedy et al. (2015); Li et al. (2019); Zhang et al. (2020); Gao et al. (2019), we devise the parallel branches that could complement long-range dependencies of a network as shown in Fig. 3. To generate the spatial bias, we encode long-range dependencies by compressing feature maps in channel and spatial directions. Then, we extend it to concatenate the spatial bias map and the convolutional feature map in the channel direction. Global knowledge from spatial bias is aggregated with the local features of the convolutional feature map, so the network learns both

local and global information. As shown in the Fig. 2, the spatial bias learns a wider region while convolution focuses on the local features in an image. Therefore, the CNN with the spatial bias learns richer information through our concatenated feature map. The following section introduces the specific process of aggregating convolutional feature map and spatial bias.

Generating Spatial Bias Map. Let the input feature map X of a convolution layer be $X \in \mathbb{R}^{H \times W \times C}$. On this feature map, we compress it in the channel and spatial direction. Specifically, we use 1×1 convolution for channel reduction, where the feature map has a C' channel. Then, we adopt an average pooling layer for the spatial compression that yields $P \in \mathbb{R}^{H' \times W' \times C'}$. We get a transformed feature map by flattening each channel of the feature map P into a 1D vector, $Q \in \mathbb{R}^{1 \times C' \times H'W'}$.

To aggregate global knowledge on this transformed feature map, we exploit a $1 \times N$ convolution in the channel dimension where the N is larger than 1 so that we encode the inter-channel relationship global knowledge. The spatial bias map is then upsampled to the same size as the convolutional feature map using bilinear interpolation. The upsampled spatial bias map is concatenated with the convolutional feature maps as Eq. 1.

$$\text{Output} = \text{ReLU}(\text{BN}[\text{Conv}(X), SB]), \quad (1)$$

where X denotes an input feature map, $\text{Conv}()$ and SB denote a standard convolution and a spatial bias, and $[,]$ is the concatenate operation. After concatenation, the resultant feature map is sent through batch normalization and nonlinear activation layers.

Convolution with Spatial Bias. In general, naive convolution employs modest kernel sizes (*e.g.*, 3×3). To compensate the limited kernel size, the self-attention module is added after specific convolution layer to learn the global knowledge. In other words, the heavy self-attention operation is independently applied which increase parameters and computational budget extensively. The proposed spatial bias and convolutional features are complementary to each other. Spatial bias extracts the information of long range-dependency, which complement the existing short range-dependency of convolutional operation. The proposed spatially biased convolution need only minimal overhead of convolution operation due to our self-attention free approach. We aim to learn both of the local and global knowledge in convolution layer directly without additional module.

Complexity of Spatial Bias. In this paragraph, we discuss about the complexity of the proposed spatial bias in comparison with the self-attention operation. Suppose input size is defined as $X \in \mathbb{R}^{H \times W \times C}$, the self-attention mechanism has a computational complexity of $O((HW)^2C) \approx O((HW)^2)$, because self-attention operation applies three projections to obtain query, keys and values and computes dot products between query and keys. On the other hand, the proposed spatial bias reduces the feature map size X by a fixed constant. Therefore, the complexity of spatial bias is $O(H'W'sf) \approx O(H'W'f)$, where s and f denote the kernel size, number of filters. Since the number of filters is the same as $H'W'$, the spatial bias has the complexity of $O((H'W')^2)$. The reduced height H' and width W' are fixed constant value, so the computational complexity is ideally $O(1)$.

Therefore, we get very fast and lightweight operation that can be inserted into any convolutional layers. In the experiment section, we show its effectiveness with regard to the throughput, parameters, and computational overhead as well as performance improvement of CNNs.

On the Comparison with Squeeze and Excitation. The general channel attention method (*i.e.*, SENet, SKNet) Hu et al. (2018); Li et al. (2019) captures the channel-wise summarized information of the feature map and then learns the relationship between the channels to adjust the feature map, so the spatial correlation is not learned. On the other hand, the proposed spatial bias extracts the spatial-wise compressed information and then expands it toward the channel. That is, dependence on the spatial-channel direction can be aggregated at once with only a general convolution operation. In addition, while SE-like method refine the channel importance of the output feature map of the convolution layer, the proposed method learn different information in that it learns the channel and spatial information together in the convolution process directly by adding a bias to the feature map. Therefore, as shown in Table 8, the proposed spatial bias is more efficient than SE-like methods and additionally improve the performance of backbone when combining them with our spatial bias.

4. Experiments

In this section, we first perform ablation studies on the proposed spatial bias, then compare it with the conventional non-local neural networks with self-attention operation. We, then, provide the experimental analysis

Table 1: Experimental result on CIFAR-100 through the proposed Spatial Bias(**3-bias channels**). Here, $s_{\#}$ denotes the stage index of the ResNet architecture after the stem cell.

Network	Stage	Param.	MFLOPs	Top-1 Error (%)	Top-5 Error (%)
ResNet-38	-	0.43M	62.2	23.98 ± 0.23	5.73 ± 0.20
	s_1	0.45M	63.6	23.86 ± 0.08	5.60 ± 0.13
	s_1, s_2	0.46M	64.2	22.44 ± 0.10	4.99 ± 0.05
	s_1, s_2, s_3	0.48M	64.6	22.46 ± 0.25	5.21 ± 0.30
ResNet-65	-	0.71M	103.3	21.87 ± 0.35	5.33 ± 0.18
	s_1	0.74M	105.8	20.81 ± 0.41	4.73 ± 0.09
	s_1, s_2	0.77M	106.9	20.77 ± 0.14	4.67 ± 0.29
	s_1, s_2, s_3	0.80M	107.5	20.37 ± 0.20	4.61 ± 0.09
ResNet-110	-	1.17M	171.9	20.59 ± 0.38	4.96 ± 0.10
	s_1	1.22M	176.1	19.97 ± 0.08	4.55 ± 0.06
	s_1, s_2	1.28M	178.0	19.42 ± 0.20	4.38 ± 0.06
	s_1, s_2, s_3	1.34M	179.1	19.65 ± 0.68	4.80 ± 0.21

Table 2: Experimental result on the comparison of insertion position in a bottleneck. We add spatial bias in parallel after **Conv1** or **Conv2**. The out channels of **Conv2** is reduced so that the spatial bias after **Conv2** has less parameters.

Position	Stage	Top-1 Error (%)	Param.
Conv1	s_1, s_2	20.57	0.78M
	s_1, s_2, s_3	20.56	0.83M
Conv2	s_1, s_2	20.77	0.77M
	s_1, s_2, s_3	20.37	0.80M

of OOD and shape bias to support the effectiveness of the spatial bias. Finally, we show the experimental result on the object detection and semantic segmentation tasks.

4.1. Experimental result on CIFAR-100

Setup. We report mean accuracy of three experiments using the CIFAR-100 dataset on the proposed method with different random seed for reliable comparison. We set the training recipe with reference to Yun et al. (2019). We use 32×32 image size 64 samples per mini-batch with 300 epochs. We initially set the learning rate as 0.25 and decayed it by a factor of 0.1 after each 75 epochs.

Table 3: Various size for Spatial bias on CIFAR-100 (*i.e.*, SB_6 denotes compression size of 6).

Network	Param.	Top-1 Error (%)	Throughput (sample/sec)
ResNet-65	0.71M	21.87	12,816
SB_6	0.77M	20.77	10,276
SB_{10}	1.13M	20.48	10,221
SB_{14}	2.33M	20.84	10,267
SB_{16}	3.47M	20.75	10,467

Position of Spatial Bias. Table 1 summarizes the performance comparison of spatial bias positions in ResNet stages. Since the spatial bias compresses the spatial resolution to the fixed size (*i.e.*, 6 for CIFAR-100, 10 for ImageNet, Table 3), the overhead of parameters and computational budget is very small regardless of the stages. When we apply the spatial bias to stage1 (s_1) and stage2 (s_2), the performance of ResNet backbone is improved considerably. However, in the last stage (s_3), there is no performance improvement with the spatial bias. We assume that the spatial size of the last stage is too small so that the global knowledge is disappeared.

Further, we validate the position of the spatial bias in a residual bottleneck. Table 2 shows the performance comparison in the insertion position of spatial bias after **Conv1** or **Conv2** in a bottleneck. We confirm that the spatial bias after **Conv2** reduce the parameters while achieving similar performance compared to that of **Conv1**.

Number of Spatial Bias Channels. We compare the performance on the number of spatial bias channels. The channel of the spatial bias represents its importance in the concatenated output, and thus the more channels are used, the more global knowledge will be learned from the spatial bias. As shown in Table 4, we confirm that the performance and overhead are the optimal when three channels were used (**Bias-3** and **Bias-4**), but the performance is degraded beyond that (*i.e.*, **Bias-5** and **Bias-6**). It is inferred that if too much spatial bias is used, the convolution features are damaged which cause the performance deteriorates of entire networks.

Component analysis. We perform an ablation study on spatial bias component analysis. **Add** in Table 5 represent that the spatial bias is added to the feature map. In addition, the average pooling layer is replaced by the maximum pooling layer (**Maxpool**). Lastly, the global context is aggregated

Table 4: Experimental results on the number of spatial bias channels in CIFAR-100. Bias-# indicates the number of channels. We use ResNet-65 as the backbone networks. We add the spatial bias to the stage 1 and 2.

Method	Param.	Top-1 Error (%)
Bias-0	0.71M	21.87
Bias-1	0.76M	20.91
Bias-2	0.77M	20.99
Bias-3	0.77M	20.77
Bias-4	0.77M	20.60
Bias-5	0.77M	21.06
Bias-6	0.77M	20.82

Table 5: Component analysis for Spatial bias on CIFAR-100.

Network	Param.	Top-1 Error (%)
Add	0.76M	20.93
Maxpool	0.77M	21.03
Pool only	0.71M	22.34
<i>SB</i> ₆	0.77M	20.77

with only the average layer(**Pool only**) for performance verification on the key operations.

4.2. Experimental results on ImageNet

Setup. In this section, we present experimental result on ImageNet-1k, which includes more than 1.2 million images with 1,000 class labels. We validate our performance on ImageNet dataset using two training recipes. First, we use the training recipe of Wightman et al. (2021) to demonstrates the effectiveness of the proposed spatial bias. In this recipe, the training mini-batch size is set to 512 with 100 epochs using 160×160 input size. We initialize the learning rate to 8e-3 with the cosine learning rate scheduler. For optimization, we use a lamb You et al. (2019) that is suitable for training of large batch size. Second, we follow NLNet Chi et al. (2020)'s training recipe for fair comparison with state-of-the-art non-local neural networks. Specifically, we exploit the cropped input image as 224×224 size. The initial learning rate of 0.1 is reduced by 0.1 after 30, 60, and 80 epochs. We use 256 batch size

Table 6: Experimental results on the position of spatial bias in ImageNet-1K.

Method	Stage	Param.	Top-1 (%)	Top-5 (%)
Bias-0	-	25.56M	76.42	92.87
Bias-3	s_1, s_3	25.86M	76.68	93.13
	s_2, s_3	25.89M	77.00	93.00
	s_1, s_2, s_3	25.99M	77.11	93.19
	s_1, s_2, s_3, s_4	26.02M	76.70	93.08

Table 7: Experimental result on the number of spatial bias channels in ImageNet-1K.

Bias-#	1	2	3	4	5
Top-1 (%)	76.70	76.84	77.00	77.00	76.66
Param.	25.87M	25.88M	25.89M	25.90M	25.91M

and stochastic gradient descent(SGD) optimizer.

Result. We perform experiments with the position of the spatial bias and its channel size. We compare the performance of spatial bias on which stage of ResNet backbone. As shown in Table 6, the spatial bias has the best performance when adding it from stage 1 to 3 s_1, s_2, s_3 . This result is the same as that of CIFAR-100 result where the spatial bias does not work on the small spatial size of the last stage. In addition, when the spatial bias is not used at the first stage s_1 , the performance increase is insignificant. This result means that, as in previous studies Wang et al. (2018); Chi et al. (2020), global knowledge exists a lot in the earlier layer with high resolution, and thus the effect of spatial bias is greater.

Table 7 shows the performance comparison in the number of spatial bias channels in ImageNet dataset. When 3~4 channels of the spatial bias are added, the performance is improved by +0.58% compared to the baseline, and wider channels 5~6 degrade the performance. This result also confirm that the optimal channels should be used to avoid the damage of convolution feature map.

In Table 9, we compare the performance of spatial bias and conventional non-local neural networks Wang et al. (2018); Cao et al. (2019); Chi et al. (2020). Our spatial bias need a minimal parameter overhead compared to

⁴Results are from Chi et al. (2020)

Table 8: Experimental results on the standard attention operation. Unlike channel attention operation, spatial bias learns channel and spatial-wise dependencies to readjust the feature map. In addition, our spatial bias is lighter than channel attention operation, and has faster inference speed. Therefore, the channel attention network to which spatial bias is added improves performance with only a very small additional budget.

Network	Param.	GFLOPs	Top-1 (%)	Top-5 (%)	Throughput (sample/sec)
ResNet-50	25.56M	4.11	76.05	92.80	2042
SRM-ResNet-50 ⁴	25.62M	4.15($\Delta 0.04$)	77.10	-	1243($\Delta 799$)
GE-ResNet-50 ⁴	31.12M	4.14($\Delta 0.03$)	76.80	-	1365($\Delta 677$)
SE-ResNet-50	28.09M	4.14($\Delta 0.03$)	76.84	93.45	1787($\Delta 255$)
SK-ResNet-50	27.49M	4.47($\Delta 0.36$)	77.56	93.62	1557($\Delta 485$)
SB-ResNet-50	25.99M	4.13($\Delta 0.02$)	76.86	93.33	1836($\Delta 206$)
SB-SE-ResNet-50	28.52M	4.16($\Delta 0.05$)	77.10	93.59	1626($\Delta 416$)
SB-SK-ResNet-50	27.94M	4.49($\Delta 0.38$)	77.93	93.54	1440($\Delta 602$)

NLNet Wang et al. (2018) so that ours is faster than them by 3.3 times. Compared with improved version of non-local neural networks (*i.e.*, GCNet and BAT) Cao et al. (2019); Chi et al. (2020), the computational budget of the spatial bias is much cheaper while achieving comparative performance. In particular, existing non-local methods (NLNet, GCNet, and BAT) apply the self-attention module in only specific layers due to the heavy design, but the proposed spatial bias can be used to all layers with minimal overhead. Therefore, our spatial bias is combined with the existing non-local methods in a network to further improve its performance. We also proceed with the comparison by visualizing the attention map of spatial bias and other non-local neural networks. As shown in Fig. 4, the proposed spatial bias is simple yet straightforward, but the visualization results of our method are comparable to complex self-attention-based non-local neural networks.

4.3. Compare with Compressed Self-attention

We conduct further experiments on compressed non-local neural networks. We applied NLNet-50 by compressing the features to 10×10 with the same average pooling used in our spatial bias. As shown in Table 10, we confirms that the compressed NLNet-50 has little gain in parameters and latency, but its performance deteriorate.

4.4. OOD distortion Robustness and Shape bias

In this section, we verify the performance of the proposed spatial bias on the out-of-distribution data set for measuring a statistically different dis-

Table 9: Experimental result on the comparison with state-of-the-art (SoTA) non-local neural networks. We compare the proposed attention-free spatial bias method with the self-attention based non-local neural networks. Our spatial bias (SB) improve the performance with very few additional resources compared to the SoTA methods. Additionally, due to our lightweight architecture, SB further improve the networks when combining with self-attention based non-local methods.

Network	Param.	GFLOPs	Top-1 (%)	Top-5 (%)	Throughput (sample/sec)
ResNet-50	25.56M	4.11	76.05	92.80	2042
NLNet-50	32.92M	7.66(Δ 3.55)	77.25	93.66	1353(Δ 689)
GCet-50 _{all}	28.08M	4.12(Δ 0.01)	76.93	93.25	1719(Δ 323)
BAT-50	30.23M	5.41(Δ 1.30)	77.78	94.01	1232(Δ 810)
SB-ResNet-50	25.99M	4.13(Δ 0.02)	76.86	93.33	1836(Δ 206)
SB-NLNet-50	33.35M	7.68(Δ 3.57)	77.59	93.74	1276(Δ 766)
SB-GCNet-50 _{all}	28.51M	4.14(Δ 0.03)	77.00	93.27	1613(Δ 429)
SB-BAT-50	30.66M	5.43(Δ 1.32)	78.06	93.97	1153(Δ 889)

Table 10: Comparison between non-local neural networks and spatial bias on ImageNet-1K.

Network	Param.	Top-1 (%)	Latency (step/sec)
NLNet-50	32.9M	77.2	Δ689
Reduced NLN	32.9M	77.0	Δ630
SB (Ours)	26.0M	76.9	Δ206

tribution from the training data. We conduct an OOD distortion robustness experiment on a total of 17 test datasets which have statistically different distributions. It includes five datasets(sketches Wang et al. (2019), edge-filtered images, silhouettes, texture-shape cue conflict, and stylized images Geirhos et al. (2018a)) and 12 test datasets Geirhos et al. (2018b). We compare the OOD robustness of the spatial bias and non-local neural networks using two different metrics (accuracy difference², observed consistency³). In Table 11, the proposed spatial bias is more robust to OOD datasets compared to the conventional non-local neural networks. This result prove that the proposed spatial bias works well regardless of the data domain.

²Compare machine and human accuracy in various ood tests

³The fraction of human and models making the same choices (right or wrong) Geirhos et al. (2020)

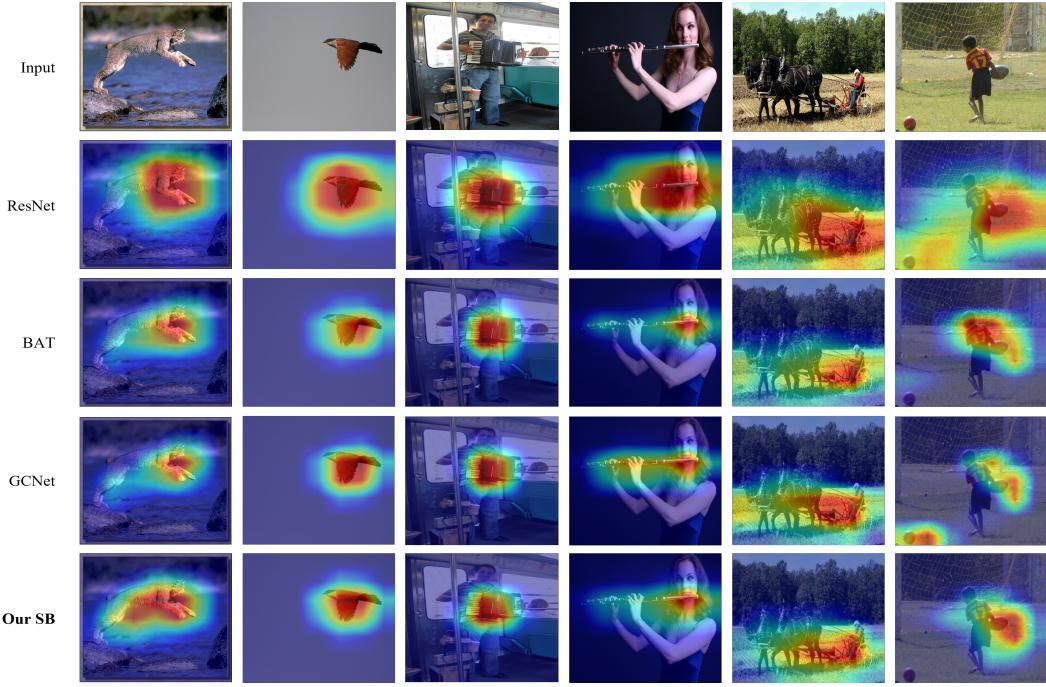


Figure 4: Grad-CAM Selvaraju et al. (2017) visualization of our spatial bias (SB) and non-local methods. Our spatial bias focuses more on the discriminant parts of an object.

4.5. Object Detection

In this subsection, we validate the performance of spatial bias on object detection task. In this experiment, we use Faster R-CNN Ren et al. (2015) and Cascade R-CNN Cai and Vasconcelos (2018) with FPN Lin et al. (2017) using 118k training and 5k verification images from the MS COCO-2017 dataset Lin et al. (2014). We use ResNet-50 as a backbone models previously trained on ImageNet dataset. By following the standard protocol Chen et al. (2019), we use a $1 \times$ learning rate schedule with 12 epochs. We exploit the SGD optimizer with $1e-4$ weight decay value and 0.9 momentum, initial learning rate as 0.02. Networks are trained on two A5000 GPUs with 8 samples per GPU. We reduce the width of the image to 800 and keep the height below 1,333 pixels. As shown in Table 12 , the networks with our spatial bias improve the performance of detection model for all metrics.

Table 11: Experimental results on OOD datasets. We compare the OOD robustness using three metrics. Spatial bias shows better OOD robustness compared to non-local neural networks.

Network	Acc diff. ↓	Obs.consistency ↑
BAT-50	0.069	0.677
SBNet-50	0.078	0.668
GCNet-50	0.081	0.668
NLNet-50	0.086	0.661
ResNet-50	0.087	0.657

Table 12: Experimental results on object detection using MS-COCO dataset.

Method	Backbone	AP	AP ₅₀	AP ₇₅	AP _S	AP _M	AP _L
Faster-RCNN	ResNet-50	39.0	60.3	42.4	23.0	43.2	50.0
	ResNet-50 + ours	40.0	61.5	43.7	24.0	44.1	51.6
Cascade-RCNN	ResNet-50	41.9	60.5	45.7	24.2	45.8	54.8
	ResNet-50 + ours	42.8	61.9	46.8	25.2	46.2	55.5

4.6. Semantic Segmentation

We perform the evaluation of semantic segmentation task using the ADE20k dataset Zhou et al. (2019). FPN Lin et al. (2017) architecture is utilized for the baseline model to which the spatial bias is applied⁵. Segmentation networks are trained on two GPUs with 14 samples per GPU for 40K iterations. Same as the detection networks, we use ResNet-50 backbone model trained on ImageNet with input 512×512 input image size. We employ the AdamW Loshchilov and Hutter (2017) as the optimization algorithm and set the initial learning rate as 2×10^{-4} and a weight decay as 10^{-4} with polynomial learning rate decay. As shown in Table 13, networks with our spatial bias outperform baseline networks by $+1.27aAcc$, $+2.16mIoU$, $+3.31mAcc$.

5. Conclusion

In this paper, we propose the spatial bias that learn global knowledge with fast and lightweight architecture. The proposed method adds only a

⁵We adopt the implementation of FPN model from Contributors (2020).

Table 13: Experimental result on semantic segmentation using ADE20K.

Method	Backbone	<i>aAcc</i>	<i>mIoU</i>	<i>mAcc</i>	<i>FPS(img/s)</i>
FPN	ResNet-50	77.82	38.04	48.5	40.76
	ResNet-50 + ours	79.09	40.20	51.81	32.40

few additional spatial bias channels to a convolutional feature map so that the convolution layer itself learns global knowledge with the self-attention operation. That is, the spatial bias is be a kind of non-local method that allows convolution to learn long-range dependency. Spatial bias generates much less parameter, FLOPs, and the throughput overhead than existing non-local methods Wang et al. (2018); Chi et al. (2020); Huang et al. (2019); Chen et al. (2018). Our design choice is simple yet straightforward. We assume this is the advantage of being applicable to various network architectures. We argue that the computational cost of the existing non-local neural networks with self-attention operation has increased considerably by using rather complex design choice. Also, the proposed spatial bias can be used together with existing self-attention based non-local methods. We believe that our new approach without self-attention based non local neural networks will inspire future studies.

References

- Jonathan Baxter. A model of inductive bias learning. *Journal of Artificial Intelligence Research*, 12:149–198, 2000.
- Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.
- Yue Cao, Jiarui Xu, Stephen Lin, Fangyun Wei, and Han Hu. Gcnet: Non-local networks meet squeeze-excitation networks and beyond. In *IEEE International Conference on Computer Vision Workshops*, 2019.
- 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*, 2019.

Yunpeng Chen, Yannis Kalantidis, Jianshu Li, Shuicheng Yan, and Jiashi Feng. A[^] 2-nets: Double attention networks. *Neural Information Processing Systems*, 31, 2018.

Lu Chi, Zehuan Yuan, Yadong Mu, and Changhu Wang. Non-local neural networks with grouped bilinear attentional transforms. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

Yooshin Cho, Youngsoo Kim, Hanbyel Cho, Jaesung Ahn, Hyeong Gwon Hong, and Junmo Kim. Rethinking efficacy of softmax for lightweight non-local neural networks. In *IEEE International Conference on Image Processing*. IEEE, 2022.

MMSegmentation Contributors. Mmsegmentation: Openmmlab semantic segmentation toolbox and benchmark, 2020.

Saisai Ding, Zhiyang Gao, Jun Wang, Minhua Lu, and Jun Shi. Fractal graph convolutional network with mlp-mixer based multi-path feature fusion for classification of histopathological images. *Expert Systems with Applications*, 212:118793, 2023.

Sheng Fang, Kaiyu Li, and Zhe Li. Salient positions based attention network for image classification. *Arxiv*, 2021.

Dandan Gao and Dengwen Zhou. A very lightweight and efficient image super-resolution network. *Expert Systems with Applications*, 213:118898, 2023.

Shang-Hua Gao, Ming-Ming Cheng, Kai Zhao, Xin-Yu Zhang, Ming-Hsuan Yang, and Philip Torr. Res2net: A new multi-scale backbone architecture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(2): 652–662, 2019.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *Arxiv*, 2018a.

Robert Geirhos, Carlos RM Temme, Jonas Rauber, Heiko H Schütt, Matthias Bethge, and Felix A Wichmann. Generalisation in humans and deep neural networks. *Neural Information Processing Systems*, 31, 2018b.

- Robert Geirhos, Kristof Meding, and Felix A Wichmann. Beyond accuracy: quantifying trial-by-trial behaviour of cnns and humans by measuring error consistency. *Neural Information Processing Systems*, 33:13890–13902, 2020.
- Zhen Guo, Caihong Mu, and Yi Liu. A multi-branch network based on weight sharing and attention mechanism for hyperspectral image classification. In *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, pages 5370–5373. IEEE, 2021.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.
- Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.
- Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. Ccnet: Criss-cross attention for semantic segmentation. In *IEEE International Conference on Computer Vision*, 2019.
- Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Neural Information Processing Systems*, 25, 2012.
- Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
- Guandong Li and Chunju Zhang. Faster hyperspectral image classification based on selective kernel mechanism using deep convolutional networks. *Arxiv*, 2022.
- Xiang Li, Wenhui Wang, Xiaolin Hu, and Jian Yang. Selective kernel networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.
- 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*. Springer, 2014.

Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *Arxiv*, 2017.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Neural Information Processing Systems*, 28, 2015.

Joao Paulo Schwarz Schuler, Santiago Romani Also, Domenec Puig, Hatem Rashwan, and Mohamed Abdel-Nasser. An enhanced scheme for reducing the complexity of pointwise convolutions in cnns for image classification based on interleaved grouped filters without divisibility constraints. *Entropy*, 24(9):1264, 2022.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *IEEE International Conference on Computer Vision*, 2017.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *Arxiv*, 2014.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. *Neural Information Processing Systems*, 32, 2019.

Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

Ross Wightman, Hugo Touvron, and Hervé Jégou. Resnet strikes back: An improved training procedure in timm. *Arxiv*, 2021.

- Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.
- Zhihuai Xie, Shaojie Wang, Wentian Zhao, and Zhenhua Guo. A robust context attention network for human hand detection. *Expert Systems with Applications*, 208:118132, 2022.
- Huaiqian You, Yue Yu, Marta D’Elia, Tian Gao, and Stewart Silling. Nonlocal kernel network (nkn): a stable and resolution-independent deep neural network. *Arxiv*, 2022.
- Yang You, Jing Li, Sashank Reddi, Jonathan Hseu, Sanjiv Kumar, Srinadh Bhojanapalli, Xiaodan Song, James Demmel, Kurt Keutzer, and Cho-Jui Hsieh. Large batch optimization for deep learning: Training bert in 76 minutes. *Arxiv*, 2019.
- Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *IEEE International Conference on Computer Vision*, 2019.
- Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *Arxiv*, 2016.
- Hang Zhang, Chongruo Wu, Zhongyue Zhang, Yi Zhu, Haibin Lin, Zhi Zhang, Yue Sun, Tong He, Jonas Mueller, R Manmatha, et al. Resnest: Split-attention networks. *Arxiv*, 2020.
- Richard Zhang. Making convolutional networks shift-invariant again. In *International Conference on Machine Learning*. PMLR, 2019.
- Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127(3):302–321, 2019.