

VanillaNet: the Power of Minimalism in Deep Learning

Hanting Chen¹, Yunhe Wang^{1*}, Jianyuan Guo¹, Dacheng Tao²

¹ Huawei Noah's Ark Lab. ² School of Computer Science, University of Sydney.
 {chenhanting, yunhe.wang, jianyuan.guo}@huawei.com, dacheng.tao@sydney.edu.au

Abstract

At the heart of foundation models is the philosophy of "more is different", exemplified by the astonishing success in computer vision and natural language processing. However, the challenges of optimization and inherent complexity of transformer models call for a paradigm shift towards simplicity. In this study, we introduce VanillaNet, a neural network architecture that embraces elegance in design. By avoiding high depth, shortcuts, and intricate operations like self-attention, VanillaNet is refreshingly concise yet remarkably powerful. Each layer is carefully crafted to be compact and straightforward, with nonlinear activation functions pruned after training to restore the original architecture. VanillaNet overcomes the challenges of inherent complexity, making it ideal for resource-constrained environments. Its easy-to-understand and highly simplified architecture opens new possibilities for efficient deployment. Extensive experimentation demonstrates that VanillaNet delivers performance on par with renowned deep neural networks and vision transformers, showcasing the power of minimalism in deep learning. This visionary journey of VanillaNet has significant potential to redefine the landscape and challenge the status quo of foundation model, setting a new path for elegant and effective model design. Pre-trained models and codes are available at <https://github.com/huawei-noah/VanillaNet> and <https://gitee.com/mindspore/models/tree/master/research/cv/vanillanet>.

1 Introduction

Over the past few decades, artificial neural networks have made remarkable progress, driven by the idea that increasing network complexity leads to improved performance. These networks, which consist of numerous layers with a large number of neurons or transformer blocks [43, 31], are capable of performing a variety of human-like tasks, such as face recognition [25], speech recognition [8], object detection [38], natural language processing [43], and content generation [2]. The impressive computational power of modern hardware allows neural networks to complete these tasks with both high accuracy and efficiency. As a result, AI-embedded devices are becoming increasingly prevalent in our lives, including smartphones, AI cameras, voice assistants, and autonomous vehicles.

Admittedly, one notable breakthrough in this field is the development of AlexNet [24], which consists of 12 layers and achieves state-of-the-art performance on the large-scale image recognition benchmark [7]. Building on this success, ResNet [18] introduces identity mappings through shortcut connections, enabling the training of deep neural networks with high performance across a wide range of computer vision applications, such as image classification [40], object detection [38], and semantic segmentation [33]. The incorporation of human-designed modules in these models, as well as the continued increase in network complexity, has undeniably enhanced the representational

*corresponding author

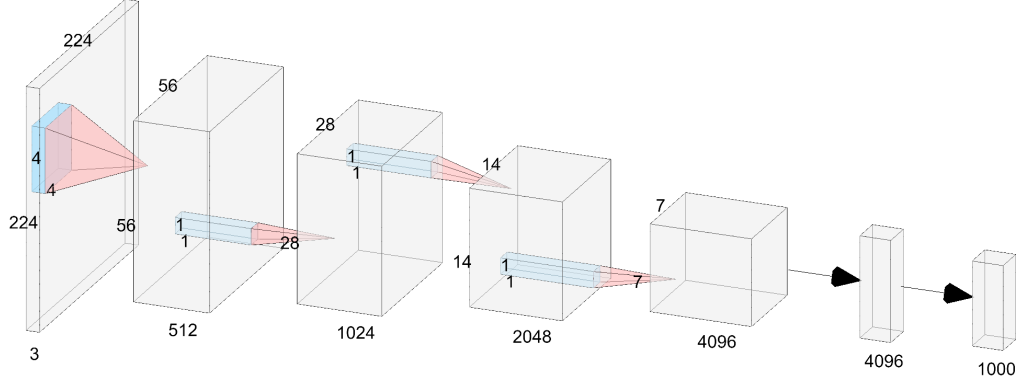


Figure 1: The architecture of VanillaNet-6 model, which consists of only 6 convolutional layers, which are very easily to be employed on any modern hardwares. The size of input features are downsampled while the channels are doubled in each stage, which borrows from the design of classical neural networks such as AlexNet [24] and VGGNet [40].

capabilities of deep neural networks, leading to a surge of research on how to train networks with more complex architectures [23, 19, 47] to achieve even higher performance.

Apart from convolutional architectures, Dosovitskiy et al.[13] have introduced the transformer architecture to image recognition tasks, demonstrating its potential for leveraging large-scale training data. Zhai et al.[53] investigated the scaling laws of vision transformer architectures, achieving an impressive 90.45% top-1 accuracy on the ImageNet dataset, which indicates that deeper transformer architectures, like convolutional networks, tend to exhibit better performance. Wang et al.[44] further proposed scaling the depth of transformers to 1,000 layers for even higher accuracy. Liu et al.[32] revisited the design space of neural networks and introduced ConvNext, achieving similar performance to state-of-the-art transformer architectures.

Although well-optimized deep and complex neural networks achieve satisfying performance, their increasing complexity poses challenges for deployment. For example, shortcut operations in ResNets consume significant off-chip memory traffic as they merge features from different layers [27]. Furthermore, complicated operations such as axial shift in AS-MLP [28] and shift window self-attention in Swin Transformer [31] require sophisticated engineering implementation, including rewriting CUDA codes.

These challenges call for a paradigm shift towards simplicity in neural network design. However, the development of ResNet has seemingly led to the abandonment of neural architectures with pure convolutional layers (without extra modules such as shortcuts). This is mainly due to the performance enhancement achieved by adding convolutional layers not meeting expectations. As discussed in [18], plain networks without shortcuts suffer from gradient vanishing, causing a 34-layer plain network to perform worse than an 18-layer one. Moreover, the performance of simpler networks like AlexNet [24] and VGGNet [40] has been largely outpaced by deep and complex networks, such as ResNets [18] and ViT [7]. Consequently, less attention has been paid to the design and optimization of neural networks with simple architectures. Addressing this issue and developing concise models with high performance would be of great value.

To this end, we propose VanillaNet, a novel neural network architecture emphasizing the elegance and simplicity of design while retaining remarkable performance in computer vision tasks. VanillaNet achieves this by eschewing excessive depth, shortcuts, and intricate operations such as self-attention, leading to a series of streamlined networks that address the issue of inherent complexity and are well-suited for resource-limited environments. To train our proposed VanillaNets, we conduct a comprehensive analysis of the challenges associated with their simplified architectures and devise a "deep training" strategy. This approach starts with several layers containing non-linear activation functions. As the training proceeds, we progressively eliminate these non-linear layers, allowing for easy merging while preserving inference speed. To augment the networks' non-linearity, we put forward an efficient, series-based activation function incorporating multiple learnable affine transfor-

mations. Applying these techniques has been demonstrated to significantly boost the performance of less complex neural networks. As illustrated in Figure 3, VanillaNet surpasses contemporary networks with elaborate architectures concerning both efficiency and accuracy, highlighting the potential of a minimalist approach in deep learning. This pioneering investigation of VanillaNet paves the way for a new direction in neural network design, challenging the established norms of foundation models and establishing a new trajectory for refined and effective model creation.

2 A Vanilla Neural Architecture

Over the past decades, researchers have reached some consensus in the basic design of neural networks. Most of the state-of-the-art image classification network architectures should consist of three parts: a stem block to transform the input images from 3 channels into multiple channels with downsampling, a main body to learn useful information, a fully connect layer for classification outputs. The main body usually has four stages, where each stage is derived by stacking same blocks. After each stage, the channels of features will expand while the height and width will decrease. Different networks utilize and stack different kinds of blocks to construct deep models.

Despite the success of existing deep networks, they utilize a large number of complex layers to extract high-level features for the following tasks. For example, the well-known ResNet [18] requires 34 or 50 layers with shortcuts for achieving over 70% top-1 accuracy on ImageNet. The base version of ViT [13] consists of 62 layers since the query, key and value in self-attention require multiple layers to calculate.

With the growing of AI chips, the bottleneck of inference speed of neural networks would not be FLOPs or parameters, since modern GPUs can easily do parallel calculation with strong computing power. In contrast, their complex designs and large depths block their speed. To this end, we propose the vanilla network, *i.e.*, VanillaNet, whose architecture is shown in Figure 1. We follow the popular design of neural network with the stem, main body and fully connect layer. Different with existing deep networks, we only employ one layer in each stage to establish an extremely simple network with as few layers as possible.

Here we show the architecture of the VanillaNet in details, which takes 6 layers as an example. For the stem, we utilize a $4 \times 4 \times 3 \times C$ convolutional layer with stride 4 following the popular settings in [18, 31, 32] to map the images with 3 channels to features with C channels. At stage 1, 2 and 3, a maxpooling layer with stride 2 is used to decrease the size and feature map and the number of channels is increased by 2. At stage 4, we do not increase the number of channels as it follows an average pooling layer. The last layer is a fully connected layer to output the classification results. The kernel size of each convolutional layer is 1×1 , since we aim to use minimal calculation cost for each layer while keep the information of feature maps. The activation function is applied after each 1×1 convolutional layer. To ease the training procedure of the network, batch normalization is also added after each layer. For the VanillaNet with different number of layers, we add blocks in each stage, which will be detailed in the supplementary material. It should be noted that the VanillaNet has no shortcut, since we empirically find adding shortcut shows little performance improvement. This also gives another benefit that the proposed architecture is extremely easy to implement since there are no branch and extra blocks such as squeeze and excitation block [22].

Although the architecture of VanillaNet is simple and relatively shallow, its weak non-linearity caused limit the performance. Therefore, we propose a series of techniques to solve the problem.

3 Training of Vanilla Networks

It is common in deep learning to enhance the performance of models by introducing stronger capacity in the training phase [4, 50]. To this end, we propose to utilize a deep training technique to bring up the ability during training in the proposed VanillaNet, since deep network have stronger non-linearity than shallow network.

3.1 Deep Training Strategy

The main idea of deep training strategy is to train two convolutional layers with an activation function instead of a single convolution layer in the beginning of training procedure. The activation function

is gradually reduce to an identity mapping with the increasing number of training epochs. At the end of training, **two convolutions can be easily merged into the one convolution to reduce the inference time**. This kind of idea is also widely used in CNNs [10, 12, 9, 11]. We then describe how to conduct this technique in detail.

For an activation function $A(x)$ (which can be the usual functions such ReLU and Tanh), we combine it with an identity mapping, which can be formulated as:

$$A'(x) = (1 - \lambda)A(x) + \lambda x, \quad (1)$$

where λ is a hyper-parameter to balance the non-linearity of the modified activation function $A'(x)$. Denote the current epoch and the number of deep training epochs as e and E , respectively. We set $\lambda = \frac{e}{E}$. Therefore, at the beginning of training ($e = 0$), $A'(x) = A(x)$, which means the network have strong non-linearity. When the training converged, we have $A'(x) = x$, which means the two convolutional layers have no activation functions in the middle. We further demonstrate how to merge these two convolutional layers.

We first convert every batch normalization layer and its preceding convolution into a single convolution. We denote $W \in \mathbb{R}^{C_{out} \times (C_{in} \times k \times k)}$, $B \in \mathbb{R}^{C_{out}}$ as the weight and bias matrices of convolutional kernel with C_{in} input channels, C_{out} output channels and kernel size k . The scale, shift, mean and variance in batch normalization are represented as $\gamma, \beta, \mu, \sigma \in \mathbb{R}^{C_{out}}$, respectively. The merged weight and bias matrices are:

$$W'_i = \frac{\gamma_i}{\sigma_i} W_i, B'_i = \frac{(B_i - \mu_i)\gamma_i}{\sigma_i} + \beta_i, \quad (2)$$

where subscript $i \in \{1, 2, \dots, C_{out}\}$ denotes the value in i -th output channels.

After merging the convolution with batch normalization, we begin to merge the two 1×1 convolutions. Denote $x \in \mathbb{R}^{C_{in} \times H \times W}$ and $y \in \mathbb{R}^{C_{out} \times H' \times W'}$ as the input and output features, the convolution can be formulated as:

$$y = W * x = W \cdot \text{im2col}(x) = W \cdot X, \quad (3)$$

where $*$ denotes the convolution operation, \cdot denotes the matrix multiplication and $X \in \mathbb{R}^{(C_{in} \times 1 \times 1) \times (H' \times W')}$ is derived from the im2col operation to transform the input into a matrix corresponding to the kernel shape. Fortunately, for 1×1 convolution, we find that the im2col operation becomes a simple reshape operation since there are no need for sliding kernels with overlap. Therefore, denote the weight matrix of two convolution layers as W^1 and W^2 , the two convolution without activation function is formulated as:

$$y = W^1 * (W^2 * x) = W^1 \cdot W^2 \cdot \text{im2col}(x) = (W^1 \cdot W^2) * X, \quad (4)$$

Therefore, 1×1 convolution can merged without increasing the inference speed.

3.2 Series Informed Activation Function

There have been proposed several different activation functions for deep neural networks, including the most popular Rectified Linear Unit (ReLU) and its variants (PReLU [17], GeLU [20] and Swish [37]). They focus on bring up the performance of deep and complex networks using different activation functions. However, as theoretically proved by the existing works [35, 14, 42], the limited power of simple and shallow network are mainly caused by the poor non-linearity, which is different with deep and complex networks and thus has not been fully investigated.

In fact, there are two ways to improve the non-linearity of a neural network: stacking the non-linear activation layers or increase the non-linearity of each activation layer, while the trend of existing networks choose the former one, which results in high latency when the parallel computation ability is excess.

One straight forward idea to improve non-linearity of activation layer is stacking. The serially stacking of activation function is the key idea of deep networks. In contrast, we turn to concurrently stacking the activation function. Denote a single activation function for input x in neural network as $A(x)$, which can be the usual functions such ReLU and Tanh. The concurrently stacking of $A(x)$ can be formulated as:

$$A_s(x) = \sum_{i=1}^n a_i A(x + b_i), \quad (5)$$

where n denotes the number of stacked activation function and a_i, b_i is the scale and bias of each activation to avoid simple accumulation. The non-linearity of the activation function can be largely enhanced by concurrently stacking. Equation 5 can be regarded as a series in mathematics, which is the operation of adding many quantities.

To further enrich the approximation ability of the series, we enable the series based function to learn the global information by varying the inputs from their neighbors, which is similar with BNET [49]. Specifically, given a input feature $x \in \mathbb{R}^{H \times W \times C}$, where H, W and C are the number of its width, height and channel, the activation function is formulated as:

$$A_s(x_{h,w,c}) = \sum_{i,j \in \{-n,n\}} a_{i,j,c} A(x_{i+h,j+w,c} + b_c), \quad (6)$$

where $h \in \{1, 2, \dots, H\}$, $w \in \{1, 2, \dots, W\}$ and $c \in \{1, 2, \dots, C\}$. It is easy to see that when $n = 0$, the series based activation function $A_s(x)$ degenerates to the plain activation function $A(x)$, which means that the proposed method can be regarded as a general extension of existing activation functions. We use ReLU as the basic activation function to construct our series since it is efficient for inference in GPUs.

We further analyze the computation complexity of the proposed activation function compared with its corresponding convolutional layer. For a convolutional layer with K kernel size, C_{in} input channels and C_{out} output channels, the computational complexity is:

$$\mathcal{O}(\text{CONV}) = H \times W \times C_{in} \times C_{out} \times k^2, \quad (7)$$

while computation cost of its series activation layer is:

$$\mathcal{O}(\text{SA}) = H \times W \times C_{in} \times n^2. \quad (8)$$

Therefore, we have:

$$\frac{\mathcal{O}(\text{CONV})}{\mathcal{O}(\text{SA})} = \frac{H \times W \times C_{in} \times C_{out} \times K^2}{H \times W \times C_{in} \times n^2} = \frac{C_{out} \times k^2}{n^2}. \quad (9)$$

Taking the 4th stage in VanillaNet-B as an example, where $C_{out} = 2048$, $k = 1$ and $n = 7$, the ratio is about 84. In conclusion, the computation cost of the proposed activation function is still much lower than the convolutional layers. More experimental complexity analysis will be shown in the following section.

4 Experiments

In this section, we conduct experiments to verify the performance of the proposed VanillaNet on large scale image classification. Ablation study is provided to investigate effectiveness of each component of the proposed VanillaNet. We also visualize the feature of VanillaNet to further study how the proposed network learns from images.

4.1 Ablation Study

In this section, we conduct ablation study to investigate the effectiveness of proposed modules, including the series activation function and the deep training technique. Besides, we analyze the influence of adding shortcuts in the proposed VanillaNet.

Influence of number of series in activation function.

In the above section, we propose a series activation function to enhance the performance of plain activation function and enable global information exchange in feature maps. Table 1 shows the performance of the proposed VanillaNet using different number of n in Equation 6. When $n = 0$, the activation function degenerate into the plain ReLU activation function. Although the inference speed of this network is higher than using the series activation function,

Table 1: Ablation study on the number of series.

n	FLOPs (B)	Latency (ms)	Top-1 (%)
0	5.83	1.96	60.53
1	5.86	1.97	74.53
2	5.91	1.99	75.62
3	5.99	2.01	76.36
4	6.10	2.18	76.43

the network can only achieve a 60.53% top-1 accuracy on the ImageNet dataset, which cannot be applied in real-world applications. It proves that the poor non-linearity of activation function results in poor performance of vanilla networks.

To this end, we propose the series activation function. When $n = 1$, the network can achieve a 74.53% accuracy, which is a huge improvement compared with 60.53%. The result demonstrate the effectiveness of the proposed activation function. When the number of n increases, the performance of the network brings up. We find that $n = 3$ is a good balance in the top-1 accuracy and latency. Therefore, we use $n = 3$ for the rest experiments. It should be noted that the FLOPs of the proposed activation function is very small compared with the original network, which is the same as the conclusion we derive in Equation 9.

Influence of deep training. As the VanillaNet is very shallow, we propose to increase the training non-linearity to bring up its performance. We then analyze the effectiveness of the proposed deep training technique. Table 2 shows the results on using deep training technique with VanillaNet-6. As a result, the original VanillaNet achieves a 75.23% top-1 accuracy, which is the baseline. By using the deep training technique, the proposed VanillaNet can achieve a 76.36% accuracy. The results demonstrate that the proposed deep training technique is useful for the shallow network.

Table 2: Ablation study on different networks.

Network	Deep train.	Series act.	Top-1 (%)
VanillaNet-6	✓	✓	59.58
			60.53
	✓	✓	75.23
			76.36
AlexNet	✓	✓	57.52
			59.09
	✓	✓	61.12
			63.59
ResNet-50	✓	✓	76.13
			76.16
	✓	✓	76.30
			76.27

Moreover, we further apply the deep training and series activation function in other networks to show the generalization ability of the two techniques. Table 2 reports the results of AlexNet and ResNet-50, which are two classical deep neural networks, on the ImageNet dataset. The original AlexNet can only achieve a 57.52% accuracy with 12 layers. By applying the proposed deep training and series activation function, the performance of AlexNet can be largely brought up by about 6%, which demonstrates that the proposed technique is highly effective for shallow networks. When it turns to ResNet-50 whose architecture are relatively complex, the performance gain is little. This results suggests the deep and complex networks already have enough non-linearity without the proposed techniques.

Influence of shortcuts. In deep neural networks, a common sense is that adding shortcut can largely ease the training procedure and improve the performance [18]. To this end, we investigate whether shortcut would benefit the performance of shallow and ximple network. We propose to use two kinds of location of shortcut, *i.e.*, shortcut after the activation function and shortcut before the activation function, which are proposed in the original ResNet [18] and PreAct-ResNet [19], respectively. Since the number of channels is large and the original convolution is with kernel size 1×1 in VanillaNet, adding a shortcut (even with 1×1 kernel size) would largely increase the FLOPs. Therefore, we use the parameter-free shortcut. It should be noted that if the stride is 2, the parameter-free shortcut will use an average pooling to decrease the size of feature maps and if the number of channel is increasing, the parameter-free shortcut utilizes padding for the extra channels following the original setting in [18].

Table 3: Ablation on adding shortcuts.

Type	Top-1 (%)
no shortcut	76.36
shortcut before act	75.92
shortcut after act	75.72

Table 3 shows the ablation study on adding shortcuts. We surprisingly find that using shortcuts, in spite of any type of shortcuts, has little improvement on the performance of the proposed VanillaNet. We suggest that the bottleneck of vanilla networks is not the identity mapping, but the weak non-linearity. The shortcut is useless for bringing up the non-linearity and may decrease non-linearity since the shortcut skips the activation function to decrease the depth of the vanilla network, therefore results in lower performance.

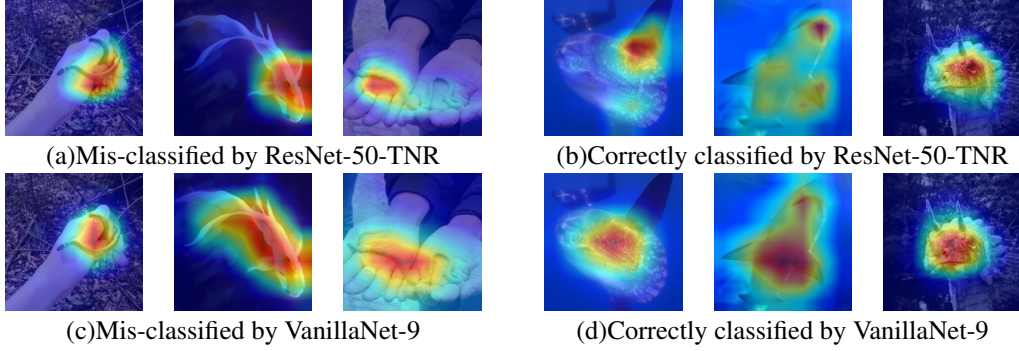


Figure 2: Visualization of attention maps of the classified samples by ResNet-50 and VanillaNet-9. We show the attention maps of their mis-classified samples and correctly classified samples for comparison.

4.2 Visualization of Attention

To have a better understanding of the proposed VanillaNet, we further visualize the features using GradCam++ [3], which utilizes a weighted combination of the positive partial derivatives of the feature maps generated by the last convolutional layer with respect to the specific class to generate a good visual explanation.

Figure 2 shows the visualization results for VanillaNet-9 and ResNets-50-TNR [45] with similar performance. The red color denotes that there are high activation in this region while the blue color denotes the weak activation for the predicted class. We can find that these two networks have different attention maps for different samples. It can be easily found that for ResNet-50, the area of active region is smaller. For the VanillaNet with only 9 depth, the active region is much larger than that of deep networks. We suggest that VanillaNet may be strong in extract all relative activations in the input images and thoroughly extract their information by using large number of parameters and FLOPs. In contrast, VanillaNet may be weak on analyzing part of the useful region since the non-linearity is relatively low.

4.3 Comparison with SOTA architectures

To illustrate the effectiveness of the proposed method, we conduct experiments on the ImageNet [7] dataset, which consists of 224×224 pixel RGB color images. The ImageNet dataset contains 1.28M training images and 50K validation images with 1000 categories. We utilize strong regularization since the proposed VanillaNet has large number of parameters in each layer to capture useful information from the images with limited non-linearity. We also report the ImageNet Real results where the labels are refined. The latency is tested on Nvidia A100 GPU.

We propose architecture for VanillaNet with different number of layers. Table 4 shows the classification results on the ImageNet dataset using different networks. We list the number of parameters, FLOPs, depth, GPU latency and accuracy for comparison. In the past decades, researchers focus on minimize the FLOPs or the latency in ARM/CPU for portable networks since they assume that the computing power in edge devices is limited. As the development of modern AI chips, several mobile devices such as driverless vehicle [26] and robots [15] are required and able to carry multiple GPUs with huge computing power for seeking real-time feedback of external inputs. Therefore, we test the GPU latency with batch size 1, which means that the AI chip has enough computing power to calculate each network. Under this situation, we find that the inference speed has little relationship with the number of FLOPs and parameters. Taking MobileNetV3-Large as an example, though it has a very low FLOPs (0.22B), its GPU latency is 7.83, which is even larger than our VanillaNet-13 with a 11.9B FLOPs. In fact, the inference speed in this setting is highly related to the complexity and number of layers. We can compare the inference speed of ShuffleNetV2x1.5 and ShuffleNetV2x2. In fact, their difference only lies in the number of channels. Therefore, although their number of parameters and FLOPs differs a lot. (0.3B v.s. 0.6B), their inference speed is nearly the same (7.23 and 7.84). We can also find in Table 4 that the straightforward architecture including

Table 4: Comparison on ImageNet. Latency is tested on Nvidia A100 GPU with batch size of 1.

Model	Params (M)	FLOPs (B)	Depth	Latency (ms)	Acc (%)	Real Acc (%)
MobileNetV3-Small [21]	2.5	0.06	48	6.65	67.67	74.33
MobileNetV3-Large [21]	5.5	0.22	48	7.83	74.04	80.01
ShuffleNetV2x1.5 [39]	3.5	0.30	51	7.23	73.00	80.19
ShuffleNetV2x2 [21]	7.4	0.58	51	7.84	76.23	82.72
RepVGG-A0 [12]	8.1	1.36	23	3.22	72.41	79.33
RepVGG-A1 [12]	12.8	2.37	23	3.24	74.46	81.02
RepVGG-B0 [12]	14.3	3.1	29	3.88	75.14	81.74
RepVGG-B3 [12]	110.9	26.2	29	4.21	80.50	86.44
ViTAE-T [48]	4.8	1.5	67	13.37	75.3	82.9
ViTAE-S [48]	23.6	5.6	116	22.13	82.0	87.0
ViTAEV2-S [55]	19.2	5.7	130	24.53	82.6	87.6
ConvNextV2-A [46]	3.7	0.55	41	6.07	76.2	82.79
ConvNextV2-F [46]	5.2	0.78	41	6.17	78.0	84.08
ConvNextV2-P [46]	9.1	1.37	41	6.29	79.7	85.60
ConvNextV2-N [46]	15.6	2.45	47	6.85	81.2	-
ConvNextV2-T [46]	28.6	4.47	59	8.40	82.5	-
ConvNextV2-B [46]	88.7	15.4	113	15.41	84.3	-
Swin-T [31]	28.3	4.5	48	10.51	81.18	86.64
Swin-S [31]	49.6	8.7	96	20.25	83.21	87.60
ResNet-18-TNR [45]	11.7	1.8	18	3.12	70.6	79.4
ResNet-34-TNR [45]	21.8	3.7	34	5.57	75.5	83.4
ResNet-50-TNR [45]	25.6	4.1	50	7.64	79.8	85.7
VanillaNet-5	15.5	5.2	5	1.61	72.49	79.66
VanillaNet-6	32.5	6.0	6	2.01	76.36	82.86
VanillaNet-7	32.8	6.9	7	2.27	77.98	84.16
VanillaNet-8	37.1	7.7	8	2.56	79.13	85.14
VanillaNet-9	41.4	8.6	9	2.91	79.87	85.66
VanillaNet-10	45.7	9.4	10	3.24	80.57	86.25
VanillaNet-11	50.0	10.3	11	3.59	81.08	86.54
VanillaNet-12	54.3	11.1	12	3.82	81.55	86.81
VanillaNet-13	58.6	11.9	13	4.26	82.05	87.15
VanillaNet-13-1.5 \times	127.8	26.5	13	7.83	82.53	87.48
VanillaNet-13-1.5 \times [†]	127.8	48.9	13	9.72	83.11	87.85

ResNet, VGGNet and our VanillaNet without extra branch and complex blocks (*e.g.*, squeeze and excitation block or densely connects) achieves the highest inference speed.

To this end, we propose the VanillaNet, which is simple and has few convolutional layers without any branch (even without shortcut). We set different number of layers in VanillaNets to construct a series of networks. As shown in Table 4, the VanillaNet-9 achieves a 79.87% accuracy with only a 2.91ms inference speed in GPU, which is over 50% faster than the ResNet-50 and ConvNextV2-P with similar performance. The surprising result demonstrate the potential of VanillaNet in real-time processing over the existing deep networks. We also scale the number of channels and the pooling size to obtain the proposed VanillaNet-13-1.5 \times [†], which achieves an 83.11% Top-1 accuracy on ImageNet, which suggests that the proposed vanilla neural network still have power to obtain such a high performance on large scale image classification task. It is suggested that we may not need deep and complex networks on image classification since scaling up VanillaNets can achieve similar performance with deep networks.

The Figure 3 shows the depth and inference speed of different architectures. The inference speed with batch size 1 is highly related to the depth of the network instead of the number of parameters, which suggest that simple and shallow networks have huge potential in real-time processing. It can be easily find that the proposed VanillaNet achieve the best speed-accuracy trade-off among all these architectures with low GPU latency, which demonstrates the superiority of the proposed VanillaNet when the computing power is sufficient.

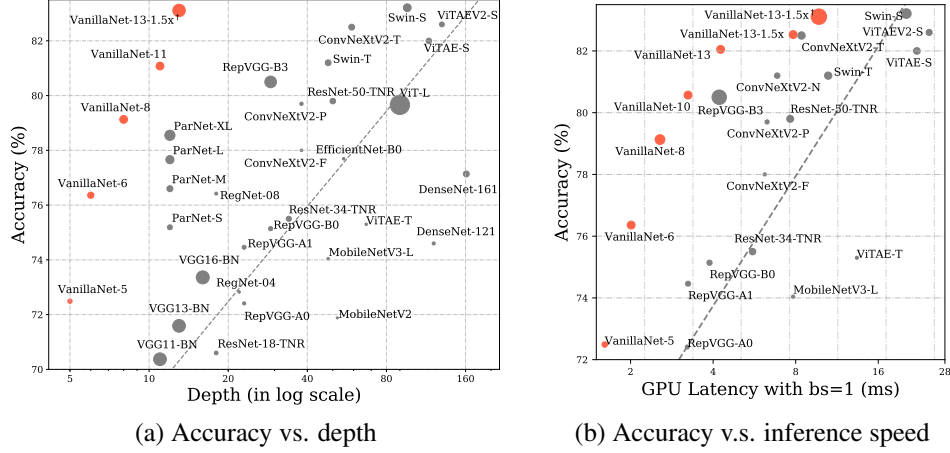


Figure 3: Top-1 Accuracy on ImageNet v.s. inference speed on Nvidia A100 GPU with batch size 1. Size of the circle is related to the depth and parameters of each architecture in (a) and (b), respectively. VanillaNet achieves comparable performance with deep neural networks while with much smaller depth and latency.

Table 5: Performance on COCO detection and segmentation. FLOPs are calculated with image size (1280, 800) on Nvidia A100 GPU.

Framework	Backbone	FLOPs	Params	FPS	AP ^b	AP ^b ₅₀	AP ^b ₇₅	AP ^m	AP ^m ₅₀	AP ^b ₇₅
RetinaNet [29]	Swin-T [31]	245G	38.5M	27.5	41.5	62.1	44.2	-	-	-
	VanillaNet-13	397G	74.6M	29.8	41.8	62.8	44.3	-	-	-
Mask RCNN [16]	Swin-T [31]	267G	47.8M	28.2	42.7	65.2	46.8	39.3	62.2	42.2
	VanillaNet-13	421G	76.3M	32.6	42.9	65.5	46.9	39.6	62.5	42.2

4.4 Experiments on COCO

To further demonstrate the effectiveness of the proposed VanillaNet on downstream tasks, we conduct evaluation in the COCO dataset [30]. We use RetinaNet [29] and Mask-RCNN [16] as the framework to evaluate the proposed method. FPS is measured on Nvidia A100 GPU.

Table 5 shows the performance of the proposed VanillaNet on COCO detection and segmentation. The proposed VanillaNet can successfully achieve similar performance with the ConvNext and the Swin backbone. Although the FLOPs and Parameters of VanillaNet is much higher than Swin and ConvNext, it has much higher FPS, which demonstrates the effectiveness of vanilla architectures on object detection and instance segmentation tasks.

5 Conclusion

This paper fully investigates the feasibility of establishing neural networks with high performance but without complex architectures such as shortcut, high depth and attention layers, which embodies the paradigm shift towards simplicity and elegance in design. We present a deep training strategy and the series activation function for VanillaNets to enhance its non-linearity during both the training and testing procedures and bring up its performance. Experimental results on large scale image classification datasets reveal that VanillaNet performs on par with well-known deep neural networks and vision transformers, thus highlighting the potential of minimalism in deep learning. We will further explore better parameter allocation for efficient VanillaNet architectures with high performance. In summary, we prove that it is possible to achieve comparable performance with the state-of-the-art deep networks and vision transformers using a very concise architecture, which will unlock the potential of vanilla convolutional network in the future.

References

- [1] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.
- [2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
- [3] Aditya Chattopadhyay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pages 839–847. IEEE, 2018.
- [4] Hanting Chen, Yunhe Wang, Chang Xu, Chao Xu, Chunjing Xu, and Tong Zhang. Universal adder neural networks. *arXiv preprint arXiv:2105.14202*, 2021.
- [5] Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*, 2020.
- [6] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 702–703, 2020.
- [7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.
- [8] Li Deng, Geoffrey Hinton, and Brian Kingsbury. New types of deep neural network learning for speech recognition and related applications: An overview. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 8599–8603. IEEE, 2013.
- [9] Xiaohan Ding, Yuchen Guo, Guiguang Ding, and Jungong Han. Acnet: Strengthening the kernel skeletons for powerful cnn via asymmetric convolution blocks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1911–1920, 2019.
- [10] Xiaohan Ding, Xiangyu Zhang, Jungong Han, and Guiguang Ding. Diverse branch block: Building a convolution as an inception-like unit. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10886–10895, 2021.
- [11] Xiaohan Ding, Xiangyu Zhang, Jungong Han, and Guiguang Ding. Scaling up your kernels to 31x31: Revisiting large kernel design in cnns. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11963–11975, 2022.
- [12] Xiaohan Ding, Xiangyu Zhang, Ningning Ma, Jungong Han, Guiguang Ding, and Jian Sun. Repvgg: Making vgg-style convnets great again. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13733–13742, 2021.
- [13] 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.
- [14] Ronen Eldan and Ohad Shamir. The power of depth for feedforward neural networks. In *Conference on learning theory*, pages 907–940. PMLR, 2016.
- [15] Shixiang Gu, Ethan Holly, Timothy Lillicrap, and Sergey Levine. Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates. In *2017 IEEE international conference on robotics and automation (ICRA)*, pages 3389–3396. IEEE, 2017.
- [16] 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.
- [17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.
- [18] 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.

- [19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 630–645. Springer, 2016.
- [20] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.
- [21] 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 *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1314–1324, 2019.
- [22] 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.
- [23] 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.
- [24] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1097–1105, 2012.
- [25] Steve Lawrence, C Lee Giles, Ah Chung Tsoi, and Andrew D Back. Face recognition: A convolutional neural-network approach. *IEEE transactions on neural networks*, 8(1):98–113, 1997.
- [26] Jesse Levinson, Jake Askeland, Jan Becker, Jennifer Dolson, David Held, Soeren Kammel, J Zico Kolter, Dirk Langer, Oliver Pink, Vaughan Pratt, et al. Towards fully autonomous driving: Systems and algorithms. In *2011 IEEE intelligent vehicles symposium (IV)*, pages 163–168. IEEE, 2011.
- [27] Guilin Li, Junlei Zhang, Yunhe Wang, Chuanjian Liu, Matthias Tan, Yunfeng Lin, Wei Zhang, Jiashi Feng, and Tong Zhang. Residual distillation: Towards portable deep neural networks without shortcuts. *Advances in Neural Information Processing Systems*, 33:8935–8946, 2020.
- [28] Dongze Lian, Zehao Yu, Xing Sun, and Shenghua Gao. As-mlp: An axial shifted mlp architecture for vision. *arXiv preprint arXiv:2107.08391*, 2021.
- [29] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.
- [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 *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.
- [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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021.
- [32] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11976–11986, 2022.
- [33] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, pages 3431–3440, 2015.
- [34] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.
- [35] Hrushikesh N Mhaskar and Tomaso Poggio. Deep vs. shallow networks: An approximation theory perspective. *Analysis and Applications*, 14(06):829–848, 2016.
- [36] Boris T Polyak and Anatoli B Juditsky. Acceleration of stochastic approximation by averaging. *SIAM journal on control and optimization*, 30(4):838–855, 1992.
- [37] Prajit Ramachandran, Barret Zoph, and Quoc V Le. Searching for activation functions. *arXiv preprint arXiv:1710.05941*, 2017.
- [38] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, pages 91–99, 2015.

- [39] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.
- [40] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- [41] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Re-thinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.
- [42] Matus Telgarsky. Benefits of depth in neural networks. In *Conference on learning theory*, pages 1517–1539. PMLR, 2016.
- [43] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- [44] Hongyu Wang, Shuming Ma, Li Dong, Shaohan Huang, Dongdong Zhang, and Furu Wei. Deepnet: Scaling transformers to 1,000 layers. *arXiv preprint arXiv:2203.00555*, 2022.
- [45] Ross Wightman, Hugo Touvron, and Hervé Jégou. Resnet strikes back: An improved training procedure in timm. *arXiv preprint arXiv:2110.00476*, 2021.
- [46] Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. Convnext v2: Co-designing and scaling convnets with masked autoencoders. *arXiv preprint arXiv:2301.00808*, 2023.
- [47] 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.
- [48] Yufei Xu, Qiming Zhang, Jing Zhang, and Dacheng Tao. Vitae: Vision transformer advanced by exploring intrinsic inductive bias. *Advances in Neural Information Processing Systems*, 2021.
- [49] Yuhui Xu, Lingxi Xie, Cihang Xie, Wenrui Dai, Jieru Mei, Siyuan Qiao, Wei Shen, Hongkai Xiong, and Alan Yuille. Bnet: Batch normalization with enhanced linear transformation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.
- [50] Jiwei Yang, Xu Shen, Jun Xing, Xinmei Tian, Houqiang Li, Bing Deng, Jianqiang Huang, and Xian-sheng Hua. Quantization networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7308–7316, 2019.
- [51] 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 preprint arXiv:1904.00962*, 2019.
- [52] 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 *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032, 2019.
- [53] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12104–12113, 2022.
- [54] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.
- [55] Qiming Zhang, Yufei Xu, Jing Zhang, and Dacheng Tao. Vitaev2: Vision transformer advanced by exploring inductive bias for image recognition and beyond. *International Journal of Computer Vision*, 2023.

A Network Architectures

The detailed architecture for VanillaNet with 7-13 layers can be found in Table 6, where each convolutional layer is followed with an activation function. For the VanillaNet-13-1.5 \times , the number of channels are multiplied with 1.5. For the VanillaNet-13-1.5 \times^{\dagger} , we further use adaptive pooling for stage 2, 3 and 4 with feature shape 40 \times 40, 20 \times 20 and 10 \times 10, respectively.

	Input	VanillaNet-5	VanillaNet-6	VanillaNet-7/8/9/10/11/12/13
stem	224×224	4×4, 512, stride 4		
stage1	56×56	[1×1, 1024]×1 MaxPool 2×2	[1×1, 1024]×1 MaxPool 2×2	[1×1, 1024]×2 MaxPool 2×2
stage2	28×28	[1×1, 2048]×1 MaxPool 2×2	[1×1, 2048]×1 MaxPool 2×2	[1×1, 2048]×1 MaxPool 2×2
stage3	14×14	[1×1, 4096]×1 MaxPool 2×2	[1×1, 4096]×1 MaxPool 2×2	[1×1, 4096]×1/2/3/4/5/6/7 MaxPool 2×2
stage4	7×7	-	[1×1, 4096]×1	[1×1, 4096]×1
classifier	7×7	AvgPool 7×7 1×1, 1000		

Table 6: Detailed architecture specifications.

B Training Details

For classification on ImageNet, we train the VanillaNets for 300 epochs utilizing the cosine learning rate decay [34]. The λ in Equ. 1 is linearly decayed from 1 to 0 on epoch 0 and 100, respectively. The training details can be found in Table 7. For the VanillaNet-11, since the training difficulty is relative large, we use the pre-trained weight from the VanillaNet-10 as its initialization. The same technique is adopted for VanillaNet-12/13.

For detection and segmentation on COCO, we use the ImageNet pre-trained weight. We train the VanillaNet-11 using the Adamw optimizer with a batch size of 32, an initial learning rate of $1.4e-4$ and an 0.8 layer wise decay.

Training Config	VanillaNet-{5/6/7/8/9/10/11/12/13}
weight init	trunc. normal (0.2)
optimizer	LAMB [51]
loss function	BCE loss
base learning rate	$3.5e-3$ {5,8-13} / $4.8e-3$ {6-7}
weight decay	0.35/0.35/0.35/0.3/0.3/0.25/0.3/0.3/0.3
optimizer momentum	$\beta_1, \beta_2=0.9, 0.999$
batch size	1024
training epochs	300
learning rate schedule	cosine decay
warmup epochs	5
warmup schedule	linear
dropout	0.05
layer-wise lr decay [5, 1]	0 {5,8-12} / 0.8 {6-7,13}
randaugment [6]	(7, 0.5)
mixup [54]	0.1/0.15/0.4/0.4/0.4/0.4/0.8/0.8/0.8
cutmix [52]	1.0
color jitter	0.4
label smoothing [41]	0.1
exp. mov. avg. (EMA) [36]	0.999996 {5-10} / 0.99992 {11-13}
test crop ratio	0.875 {5-11} / 0.95 {12-13}

Table 7: ImageNet-1K training settings.