HourNAS: Extremely Fast Neural Architecture Search Through an Hourglass Lens (Supplementary Material)

Zhaohui Yang[1,2], Yunhe Wang[2]*, Xinghao Chen[2], Jianyuan Guo[2],
Wei Zhang[2], Chao Xu[1], Chunjing Xu[2], Dacheng Tao[3], Chang Xu[3]
[1] Key Lab of Machine Perception (MOE), Dept. of Machine Intelligence, Peking University.
[2] Noah's Ark Lab, Huawei Technologies. [3] School of Computer Science, Faculty of Engineering, University of Sydney.
zhaohuiyang@pku.edu.cn; yunhe.wang@huawei.com; c.xu@sydney.edu.au

A. Examine The Vital Blocks in NAS SuperNet

The NAS search space is an extended formulation of residual network [2]. For each SuperBlock [4] in the NAS SuperNet $\mathcal{S}$, it contains several operations parallel to the identity mapping (*e.g.*, $O = 9$ for FBNet [7] and one of them is zero-mapping). NAS searches for the most appropriate operation for each block and constructs the searched architecture by stacking them. The identity mapping is served as the shortcut for some SuperBlocks. Thus, we could also divide all the SuperBlocks into "vital" and "non-vital" categories. The transformations in the vital SuperBlocks are to learn the $\mathcal{F}$, while the transformations in non-vital SuperBlocks are to learn the residual.

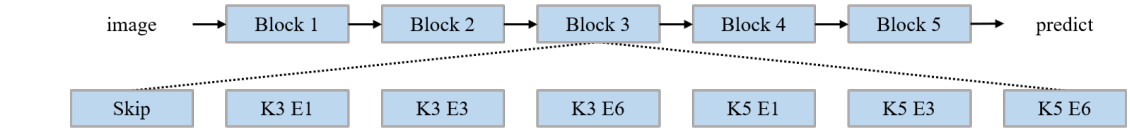
We further examine that the vital blocks play a more important role in the NAS search space by experiments. We could not train all the models (*e.g.*, 9^{22} for FBNet) in the search space. Thus we train a subset (504 models) on the ImageNet dataset to verify the hypothesis of vital blocks. The backbone is a shallower FBNet [7]. We aim to find which blocks would affect most on the final performance, in other words, which blocks are more sensitive to the final performance. If block l is to be evaluated, we fix all the other blocks with a pre-defined operation and enumerate block l with seven different operations (*e.g.*, block3 in Figure 1(a)). Then we report the mean/std. We examine six different backbones, where all the blocks are separately $ir\_k\{3,5\}\_e\{1,3,6\}$ [7][1]. All the 12 blocks in the middle are separately evaluated, and there are a total of $12 \times 7 \times 6 = 504$ models.

The statistics are shown in Figure 1. While other blocks are fixed, changing the transformations of vital blocks has the greatest impact on the final accuracy, which is reflected in greatly changing the mean or std. This phenomenon is especially obvious for experiments ir_k3_e6 and ir_k5_e6, by changing the operations of vital blocks, the averaged accuracies decrease by a large margin, and the standard deviations also increase significantly. Both analyses and experiments verify that vital blocks are more important to the final accuracy. Thus the vital blocks should be searched with a higher priority. Denoting $\mathcal{S}$ as the NAS SuperNet, We use $\mathcal{S}_{vital}$ to represent the minimal SuperNet which contains all the vital SuperBlocks.

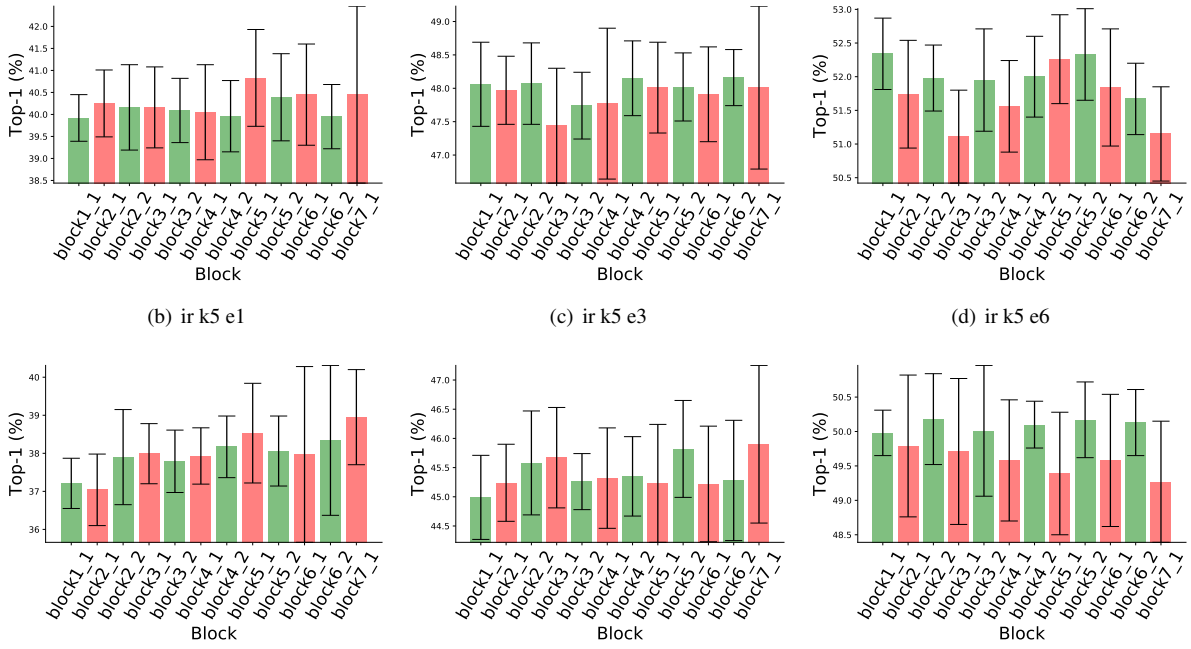
Table 1. Network definition for the sensitivity property experiments. The 'Super' denotes that this layer will be used to evaluate the mean value and standard deviation while other layers are fixed. In FBNet, each stage is constructed by 4 SuperBlocks, the first SuperBlock in each stage is vital, and the rest three of them are non-vital. In our experiments, we use 2 SuperBlocks in each stage, the first SuperBlock is vital and the next SuperBlock is non-vital.

Layer	Type	Out	Stride	vital
conv1	Conv	16	2	-
layer1_1	Super	16	1	non-vital
layer2_1	Super	24	2	vital
layer2_2	Super	24	1	non-vital
layer3_1	Super	32	2	vital
layer3_2	Super	32	1	non-vital
layer4_1	Super	64	2	vital
layer4_2	Super	64	1	non-vital
layer5_1	Super	112	1	vital
layer5_2	Super	112	1	non-vital
layer6_1	Super	184	2	vital
layer6_2	Super	184	1	non-vital
layer7_1	Super	352	1	vital
conv1	Conv	1984	1	-
avg	Avg Pool	-	-	-
fc	Fc	100	-	-

*Corresponding author.
[1]The inverted residual (ir) block [3] with kernel size k and expansion e.



(a) Sensitivity property experiments



(b) ir k5 e1 (c) ir k5 e3 (d) ir k5 e6

(e) ir k3 e1 (f) ir k3 e3 (g) ir k3 e6

Figure 1. The diagram of the block sensitivity experiments. Figure 1(b) to Figure 1(g) show the mean value and standard deviation of every block. The red histograms and the green histograms are the mean/std for the vital and non-vital blocks, respectively. A block with larger std means this block is more sensitive to the final accuracy by enumerating operations. In general, the vital blocks (red) are more sensitive to the final accuracy.

B. The Analysis of MobileNetV2

In this section, we analyze the MobileNetV2 [3] and MnasNet [5] which are utilized to measure the block importance by feature distortion in the main paper. The inverted residual block forms the basic transformation. For the blocks with stride = 1, the idendity mapping serves as the shortcut which makes the transformations less important. The MobileNetV2 is detailed in Tab. 2 and MnasNet is detailed in Tab. 9.

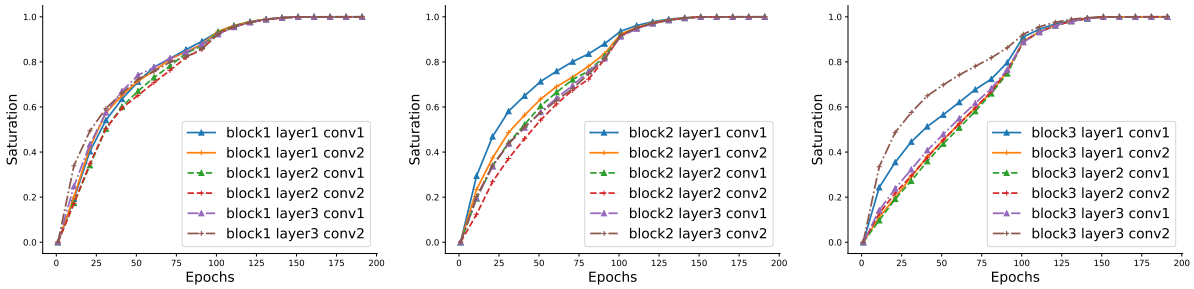
C. Other Properties of Vital Blocks

During the network optimization period, different paths have some unique and interesting properties, *e.g.*, BN bias to the shallow path [1]. In this section, we study the convergence speed of different layers in network optimization.

While training the residual network, all the paths are jointly optimized. However, the depths are different for these paths. Therefore, optimization difficulties are different. Since the minimal path is the shortest path that directly connects the input images to the ground truths. For any other path that contains more random initialized layers, we argue these paths are harder to be optimized compared to the minimal path. Thus, the minimal path converges faster. We propose a new metric to measure the convergence degree of the parameters. Networks are randomly initialized and are gradually trained until converge using the back-propagation algorithm. Each layer (transformation) learns a series of patterns and extracts the features. Denote the weights of one transformation $\mathcal{L}$ as $\mathbf{w}$, $\mathbf{w}_0$ is the randomly initialized weights, and $\mathbf{w}_t$ denotes the weights after training for t epochs. T is the maximum epoch number for training the parameter $\mathbf{w}$ until converge. We define a metric $\mathcal{C}$ to judge the convergence degree of parameter $\mathbf{w}$, and the metric is used for measuring the saturation speed of the weights. The metric $\mathcal{C}$ is defined as follows,

$$\mathcal{C}_t = 1 - \frac{\mathcal{D}(\mathbf{w}_t, \mathbf{w}_T)}{\mathcal{D}(\mathbf{w}_0, \mathbf{w}_T)}, \quad \mathcal{C}_t = \begin{cases} 0, & \text{if } t = 0, \\ 1, & \text{if } t = T. \end{cases} \tag{1}$$

where $\mathcal{D}$ is the Frobenius norm measuring the distance be-



(a) block 1 (b) block 2 (c) block 3

Figure 2. The saturation curves for weights of all the convolution layers in the ResNet20 trained on the CIFAR-10 dataset. From left to right are different layers in block1, block2, and block3, respectively. We train the network for 200 epochs. Weight decay is 1e-4. The learning rate starts from 0.1 and decays by a factor of 10 in epoch 100, 150, and 180. The 'padding' mode shortcut for downsampling blocks is used.

Table 2. Detailed architecture of MobileNetV2.

Layer	Out	Stride	Type	MnasNet-A1
Conv	32	2	Vital	conv3 × 3
layer1_1	16	1	Vital	k3e6
layer2_1	24	2	Vital	k3e6
layer2_2	24	1	Non-Vital	k3e6
layer3_1	32	2	Vital	k3e6
layer3_2	32	1	Non-Vital	k3e6
layer3_3	32	1	Non-Vital	k3e6
layer4_1	64	2	Vital	k3e6
layer4_2	64	1	Non-Vital	k3e6
layer4_3	64	1	Non-Vital	k3e6
layer4_4	64	1	Non-Vital	k3e6
layer5_1	96	1	Vital	k3e6
layer5_2	96	1	Non-Vital	k3e6
layer5_3	96	1	Non-Vital	k3e6
layer6_1	160	2	Vital	k3e6
layer6_2	160	1	Non-Vital	k3e6
layer6_3	160	1	Non-Vital	k3e6
layer7_1	320	1	Vital	k3e6
Params	-	-	-	3.4
FLOPs	-	-	-	300
Top-1 (%)	-	-	-	72.0
Top-5 (%)	-	-	-	91.0

tween two tensors. Different layers have different convergence speed, so $\mathcal{C}$ is a function related to layers.

By following the proposed saturation measurement in Eqn 1. The saturation curves are shown in Figure 2. All the weights in block1 (non-vital) converge at a similar speed. However, the vital layers show a faster saturation speed. This experiment demonstrates that different layers converge with different speeds, and the vital layers converge faster because the non-vital layers would be affected by the noisy initialization at the beginning.

D. The Diagram of Space Proposal

The figure 3 illustrates the strategy of space proposal selection. After defining the computational targets T, we optimize m different space proposals, and each space proposal is utilized for sampling architectures that satisfy the targets. The π is used for sampling space proposals. The two-level sampling strategy is used to sample architecture in every iteration. At each iteration, π is used to sample a space proposal θ_i^T and the θ_i^T is used to sample an individual architecture $A_{\theta_i^T}$.

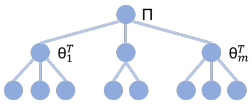


Figure 3. The diagram of space proposal selection. In this diagram, π represents the sampler for sampling the space proposals from $\Theta^T = \{\theta_1^T, \cdots, \theta_m^T\}$. The θ_i^T, $i \in \{1, \ldots, m\}$ represents the distributions for sampling the architectures $A_{\theta_i^T}$.

E. The Architectures on FBNet Search Space

HourNAS-A/B/G/I use the same search space and backbone as FBNet [7]. HourNAS searches for the kernel size, expansion ratio, and operations. The search space is defined as follows, and we also list the searched architectures in detail. It is worth noticing that we do not use bells and whistles like swish, SE modules in this experiment.

Table 3. The inverted residual block with the following settings.

Block Type	Expansion	Kernel	Group
k3g2	1	3	2
k3e1	1	3	1
k3e3	3	3	1
k3e6	6	3	1
k5g2	1	5	2
k5e1	1	5	1
k5e3	3	5	1
k5e6	6	5	1
skip	-	-	-

F. The Enlarged Search Space on FBNet Backbone

For our HourNAS-C/D/H experiment, the search space and the SuperNet are defined as below. The backbone is the same as FBNet [7] and the search space is slightly enlarged. The architectures are detailed in Table 8.

Table 4. The inverted residual block with the following settings.

Block Type	Expansion	Kernel	Group
k3e1	1	3	1
k3e3	3	3	1
k3e6	6	3	1
k5e1	1	5	1
k5e3	3	5	1
k5e6	6	5	1
k7e1	1	7	1
k7e3	3	7	1
k7e6	6	7	1
skip	-	-	-

G. The Architectures on MnasNet Search Space

For our HourNAS-E experiment, the search space and the SuperNet are defined as below, which are the same as MnasNet [5]. The architectures are detailed in Table 9.

Table 5. The inverted residual block with the following settings.

Block Type	Expansion	Kernel	Group	SE
k3e1	1	3	1	False
k3e3	3	3	1	False
k3e6	6	3	1	False
k5e1	1	5	1	False
k5e3	3	5	1	False
k5e6	6	5	1	False
k3e1se	1	3	1	True
k3e3se	3	3	1	True
k3e6se	6	3	1	True
k5e1se	1	5	1	True
k5e3se	3	5	1	True
k5e6se	6	5	1	True
skip	-	-	-	-

H. The Architectures on EfficientNet Search Space

For our HourNAS-F experiment, we use the backbone the same as EfficientNet [6]. The search space and the architectures are detailed below.

Table 6. The inverted residual block with the following settings.

Block Type	Expansion	Kernel	Group	SE
k3e1se	1	3	1	True
k3e3se	3	3	1	True
k3e6se	6	3	1	True
k5e1se	1	5	1	True
k5e3se	3	5	1	True
k5e6se	6	5	1	True
skip	-	-	-	-

Table 7. Detailed architectures of HourNAS-A/B/G/I. The backbone is same as FBNet [7].

Layer	Out	Stride	Type	SuperNet	HourNAS-A	HourNAS-B	HourNAS-G (w/o crit priori)	HourNAS-I	FBNet-Max
Conv	16	2	Vital	conv3 × 3	conv3 × 3	conv3 × 3	conv3 × 3	conv3 × 3	conv3 × 3
layer1_1	16	1	Non-Vital	Super	k3e1	k3e3	k3e3	k3e1	k5e6
layer2_1	24	2	Vital	Super	k5e6	k5e6	k5g2	k5e6	k5e6
layer2_2	24	1	Non-Vital	Super	k5g2	k3e3	k3g2	k5e1	k5e6
layer2_3	24	1	Non-Vital	Super	skip	k5g2	k3e3	k3e1	k5e6
layer2_4	24	1	Non-Vital	Super	k3g2	skip	k3e3	k3e1	k5e6
layer3_1	32	2	Vital	Super	k5e6	k5e6	k5e1	k5e6	k5e6
layer3_2	32	1	Non-Vital	Super	k3g2	k3e3	k3e1	k5e3	k5e6
layer3_3	32	1	Non-Vital	Super	k5g2	k3e3	k3e3	k5e3	k5e6
layer3_4	32	1	Non-Vital	Super	k5e1	k3e3	k3e3	k3e3	k5e6
layer4_1	64	2	Vital	Super	k5e6	k5e6	k5e3	k5e6	k5e6
layer4_2	64	1	Non-Vital	Super	k5e3	k5e6	k3e6	k5e3	k5e6
layer4_3	64	1	Non-Vital	Super	k3e1	k5e6	k3e3	k5e3	k5e6
layer4_4	64	1	Non-Vital	Super	k5e3	k3e6	k5e6	k5e6	k5e6
layer5_1	112	1	Vital	Super	k5e6	k5e6	k3e3	k5e6	k5e6
layer5_2	112	1	Non-Vital	Super	k3e3	k3e6	k3e3	k5e3	k5e6
layer5_3	112	1	Non-Vital	Super	k3e3	k3e6	k5e3	k3e1	k5e6
layer5_4	112	1	Non-Vital	Super	k3e3	k5e3	k5e3	k5e3	k5e6
layer6_1	184	2	Vital	Super	k5e6	k5e6	k3e6	k5e6	k5e6
layer6_2	184	1	Non-Vital	Super	k5e3	k3e6	k5e6	k3e3	k5e6
layer6_3	184	1	Non-Vital	Super	k5e3	k3e6	k3e6	k5e3	k5e6
layer6_4	184	1	Non-Vital	Super	k3e6	k5e6	k5e3	k5e6	k5e6
layer7_1	352	1	Vital	Super	k5e6	k5e6	k3e3	k5e6	k5e6
Params	-	-	-	-	4.8	5.5	4.7	4.8	5.7
FLOPs	-	-	-	-	298	406	297	318	583
Top-1 (%)	-	-	-	-	74.1	75.0	73.2	74.2	75.7
Top-5 (%)	-	-	-	-	91.8	92.2	91.4	91.8	92.8

Table 8. Detailed architectures of HourNAS-C/D/H. The backbone is same as FBNet [7].

Layer	Out	Stride	Type	SuperNet	HourNAS-C	HourNAS-D	HourNAS-H (w/o crit priori)
Conv	16	2	conv3 × 3	conv3 × 3	conv3 × 3	conv3 × 3	conv3 × 3
layer1_1	16	1	Vital	Super	skip	k3e3	k3e3
layer2_1	24	2	Vital	Super	k3e6	k3e6	k3e1
layer2_2	24	1	Non-Vital	Super	k3e3	k5e1	k3e3
layer2_3	24	1	Non-Vital	Super	skip	k5e1	k3e3
layer2_4	24	1	Non-Vital	Super	k3e3	k5e3	k5e1
layer3_1	32	2	Vital	Super	k5e6	k5e6	k5e3
layer3_2	32	1	Non-Vital	Super	k3e3	k5e3	k3e3
layer3_3	32	1	Non-Vital	Super	k3e3	k3e6	k3e3
layer3_4	32	1	Non-Vital	Super	k3e3	k5e1	k7e1
layer4_1	64	2	Vital	Super	k5e6	k5e6	k7e3
layer4_2	64	1	Non-Vital	Super	k3e3	k3e6	k7e3
layer4_3	64	1	Non-Vital	Super	k5e3	k3e6	k7e3
layer4_4	64	1	Non-Vital	Super	k7e1	k3e6	k7e3
layer5_1	112	1	Vital	Super	k7e6	k7e6	k7e3
layer5_2	112	1	Non-Vital	Super	k5e3	k7e3	k5e3
layer5_3	112	1	Non-Vital	Super	k5e1	k5e3	k5e3
layer5_4	112	1	Non-Vital	Super	k5e1	k5e3	k7e3
layer6_1	184	2	Vital	Super	k7e6	k7e6	k7e3
layer6_2	184	1	Non-Vital	Super	k5e3	k7e6	k3e6
layer6_3	184	1	Non-Vital	Super	k3e3	k5e6	k7e3
layer6_4	184	1	Non-Vital	Super	k3e6	k7e6	k7e3
layer7_1	352	1	Vital	Super	k7e6	k7e6	k7e6
Params	-	-	-	-	4.8	5.5	4.8
FLOPs	-	-	-	-	296	394	299
Top-1 (%)	-	-	-	-	74.1	75.3	73.5
Top-5 (%)	-	-	-	-	91.6	92.3	91.3

Table 9. Detailed architectures of MnasNet search space. The backbone is same as MnasNet [5].

Layer	Out	Stride	Type	SuperNet	HourNAS-E	MnasNet-A1
Conv	32	2	Vital	conv3 × 3	conv3 × 3	conv3 × 3
layer1_1	16	1	Vital	SepConv3 × 3	SepConv3 × 3	SepConv3×3
layer2_1	24	2	Vital	Super	k5e6	k3e6
layer2_2	24	1	Non-Vital	Super	k3e3se	k3e6
layer2_3	24	1	Non-Vital	Super	k3e1se	skip
layer2_4	24	1	Non-Vital	Super	k3e3	skip
layer3_1	40	2	Vital	Super	k5e6se	k5e3se
layer3_2	40	1	Non-Vital	Super	k5e1	k5e3se
layer3_3	40	1	Non-Vital	Super	k3e1se	k5e3se
layer3_4	40	1	Non-Vital	Super	k5e1	skip
layer4_1	80	2	Vital	Super	k5e6se	k3e6
layer4_2	80	1	Non-Vital	Super	k3e3se	k3e6
layer4_3	80	1	Non-Vital	Super	k3e3	k3e6
layer4_4	80	1	Non-Vital	Super	k3e3se	k3e6
layer5_1	112	1	Vital	Super	k5e6se	k3e6se
layer5_2	112	1	Non-Vital	Super	k3e3se	k3e6se
layer5_3	112	1	Non-Vital	Super	k3e3	skip
layer5_4	112	1	Non-Vital	Super	k3e3	skip
layer6_1	160	2	Vital	Super	k5e6se	k5e6se
layer6_2	160	1	Non-Vital	Super	k3e6	k5e6se
layer6_3	160	1	Non-Vital	Super	k5e3se	k5e6se
layer6_4	160	1	Non-Vital	Super	k5e3se	skip
layer7_1	320	1	Vital	Super	k5e6se	k3e6
Params	-	-	-	-	3.8	3.9
FLOPs	-	-	-	-	313	312
Top-1 (%)	-	-	-	-	75.7	75.7
Top-5 (%)	-	-	-	-	92.8	92.8

Table 10. Detailed architectures of EfficientNet search space. The backbone is same as EfficientNet [6].

Layer	Out	Stride	Type	SuperNet	HourNAS-F	EfficientNet-Max
Conv	32	2	Vital	conv3 × 3	conv3 × 3	conv3 × 3
layer1_1	16	1	Vital	k3e1se	k3e1se	k5e6se
layer2_1	24	2	Vital	Super	k5e6se	k5e6se
layer2_2	24	1	Non-Vital	Super	k5e1se	k5e6se
layer2_3	24	1	Non-Vital	Super	k5e1se	k5e6se
layer2_4	24	1	Non-Vital	Super	k3e1se	k5e6se
layer3_1	40	2	Vital	Super	k5e6se	k5e6se
layer3_2	40	1	Non-Vital	Super	k5e1se	k5e6se
layer3_3	40	1	Non-Vital	Super	k3e1se	k5e6se
layer3_4	40	1	Non-Vital	Super	k5e1se	k5e6se
layer4_1	80	2	Vital	Super	k5e6se	k5e6se
layer4_2	80	1	Non-Vital	Super	k3e6se	k5e6se
layer4_3	80	1	Non-Vital	Super	k3e6se	k5e6se
layer4_4	80	1	Non-Vital	Super	k3e6se	k5e6se
layer5_1	112	1	Vital	Super	k5e6se	k5e6se
layer5_2	112	1	Non-Vital	Super	k5e3se	k5e6se
layer5_3	112	1	Non-Vital	Super	k5e3se	k5e6se
layer5_4	112	1	Non-Vital	Super	k5e3se	k5e6se
layer6_1	192	2	Vital	Super	k5e6se	k5e6se
layer6_2	192	1	Non-Vital	Super	k5e6se	k5e6se
layer6_3	192	1	Non-Vital	Super	k3e6se	k5e6se
layer6_4	192	1	Non-Vital	Super	k5e6se	k5e6se
layer7_1	320	1	Vital	Super	k5e6se	k5e6se
Params	-	-	-	-	5.3	5.8
FLOPs	-	-	-	-	383	738
Top-1 (%)	-	-	-	-	77.0	78.3
Top-5 (%)	-	-	-	-	93.5	94.0

References

[1] Soham De and Samuel L. Smith. Batch normalization biases deep residual networks towards shallow paths. *arxiv*, 2019. 2

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CVPR*, 2016. 1

[3] Mark B. Sandler, Andrew G. Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. *CVPR*, 2018. 1, 2

[4] Albert Shaw, Daniel Hunter, Forrest N. Iandola, and Sammy Sidhu. Squeezenas: Fast neural architecture search for faster semantic segmentation. *ICCVW*, 2019. 1

[5] Mingxing Tan, Bo Chen, Ruoming Pang, Vijay Vasudevan, Mark Sandler, Andrew Howard, and Quoc V. Le. Mnasnet: Platform-aware neural architecture search for mobile. *CVPR*, 2018. 2, 4, 6

[6] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. *ICML*, 2019. 4, 6

[7] Bichen Wu, Xiaoliang Dai, Peizhao Zhang, Yanghan Wang, Fei Sun, Yiming Wu, Yuandong Tian, Peter Vajda, Yangqing Jia, and Kurt Keutzer. Fbnet: Hardware-aware efficient convnet design via differentiable neural architecture search. *CVPR*, 2019. 1, 3, 4, 5