

Learning to Learn Parameterized Classification Networks for Scalable Input Images

Duo Li^{1,2*}, Anbang Yao²[✉], and Qifeng Chen¹[✉]

¹ The Hong Kong University of Science and Technology

² Intel Labs China

duo.li@connect.ust.hk

anbang.yao@intel.com

cqf@ust.hk

Abstract. Convolutional Neural Networks (CNNs) do not have a predictable recognition behavior with respect to the input resolution change. This prevents the feasibility of deployment on different input image resolutions for a specific model. To achieve efficient and flexible image classification at runtime, we employ meta learners to generate convolutional weights of main networks for various input scales and maintain privatized Batch Normalization layers per scale. For improved training performance, we further utilize knowledge distillation on the fly over model predictions based on different input resolutions. The learned meta network could dynamically parameterize main networks to act on input images of arbitrary size with consistently better accuracy compared to individually trained models. Extensive experiments on the ImageNet demonstrate that our method achieves an improved accuracy-efficiency trade-off during the adaptive inference process. By switching executable input resolutions, our method could satisfy the requirement of fast adaptation in different resource-constrained environments. Code and models are available at <https://github.com/d-li14/SAN>.

Keywords: Efficient Neural Networks, Visual Classification, Scale Deviation, Meta Learning, Knowledge Distillation

1 Introduction

Although CNNs have demonstrated their dominant power in a wide array of computer vision tasks, their accuracy does not scale up and down with respect to the corresponding input resolution change. Typically, modern CNNs are constructed by stacking convolutional modules in the body, a Global Average Pooling (GAP) layer and a Fully-Connected (FC) layer in the head. When input images with different sizes are fed to a CNN model, the convolutional feature maps also vary in their size accordingly, but the subsequent GAP operation could reduce all the incoming features into a tensor with 1×1 spatial size and equal amount of channels. Thanks to the GAP layer, even trained on specific-sized input images, modern CNNs are also amenable to processing images of other sizes during the inference phase. However, the primary concern lies in that their performance is

* indicates intern at Intel Labs China. [✉] indicates corresponding authors.

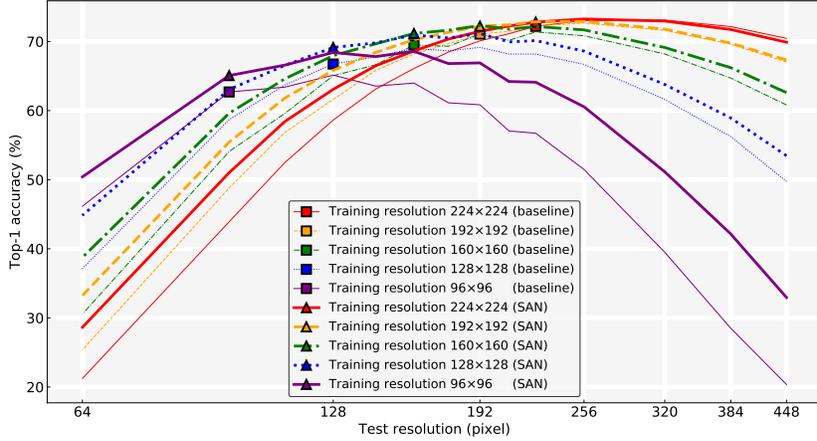


Fig. 1. Validation accuracy envelope of our proposed SAN with MobileNetV2 on ImageNet. Curves with the same color/style represent the results of models trained with the same input resolutions. The x-axis is plotted in the logarithmic scale. ▲ and ■ indicate the spots when test resolution meets the training one.

vulnerable to *scale deviation*³, exhibiting severe deterioration when evaluating images of varying sizes at the inference time, as illustrated in Fig. 1. Therefore, as done in a series of efficient network designs [13][27][12][36][20][16], in order to adapt to real-time computational demand from the aspect of input resolution, it is necessary to train a spectrum of models from scratch using input images of different resolutions. These pre-trained models are then put into a storage pool and individually reserved for future usage. Client-side devices have to retrieve pertinent models based on requirements and available resources at runtime, which will largely impede the flexibility of deployment due to inevitable downloading and offloading overheads. To flexibly handle the real-time demand on various resource-constrained application platforms, a question arises: under the premise of not sacrificing or even improving accuracy, is it possible to learn a controller to dynamically parameterize different visual classification models having a shared CNN architecture conditioned on the input resolutions at runtime?

To echo the question above, a scale-adaptive behavior is anticipated for the controller to acquire. That implies, given a CNN structure, to instantiate different main networks for each certain image resolution, the controller should have prior knowledge about switching between scaled data distributions and tactfully bridging the scaling coefficients with network parameters. We propose that appropriate data scheduling and network layer manipulations could lend this attribute to the controller. Specifically, we synthesize image patches with a set of training resolutions and employ meta networks as the controller to integrate diverse knowledge from these varying resolutions. The meta learners redistribute the gathered scale knowledge by generating convolutional kernels for multiple main networks conditioned on their assigned input resolutions respectively. Due to the tight relationship between Batch Normalization (BN) [14]

³ The concept of *scale deviation* will be discussed detailedly in Section 2.

layers and scaled activation distributions, all parameters and statistics of BN layers in each main network are maintained in a privatized manner. The main networks are collaboratively optimized following a mixed-scale training paradigm. The meta networks are optimized via collecting gradients passed through different main networks, such that information from multi-scale training flows towards the controller. Furthermore, aiming at leveraging the knowledge extracted with large resolutions to advance the performance on smaller ones, a scale distillation technique is utilized on the fly via taking the probabilistic prediction based on large resolutions as smoothed supervision signals. These aforementioned ingredients are coherently aggregated, leading to our proposed Scale Adaptive Network (SAN) framework, which is scalable by design and also generally applicable to prevailing CNN architectures.

During inference, given an input image, the learned controller could parameterize the visual classification model according to its resolution, showing consistently improved accuracy compared with the model individually trained on the corresponding resolution, as demonstrated in Fig. 1. Therefore, client devices almost merely need to reserve a single meta network and the computation graph of the backbone architecture whose parameters could be dynamically generated based on the evaluation resolution. Depending on the ability to flexibly switch target resolutions, the inference process is controllable to meet the on-device latency requirements and the accuracy expectation. Furthermore, the generated models could adapt smoothly to a wide range of input image sizes, even under the circumstances with severe problems of scale deviation. Benefiting from in-place parameterization and performance resilience on a spectrum of evaluated resolutions for each individual model, the one-model-fits-all style addresses the major obstacle of application to various scenarios.

Our main contributions can be summarized as follows:

- We employ meta networks to decide the main network parameters conditioned on its input resolution at runtime. Little research attention has been paid to this kind of meta-level information before us. We also extend the scope of knowledge distillation, based on the same image instance with different resolutions, which is also a rarely explored data-driven application scenario.
- We develop a new perspective of efficient network design by combining the aforementioned two components to permit adaptive inference on universally scaled input images, make a step forward in pushing the boundary of the resolution-accuracy spectrum and facilitating flexible deployment of visual classification models among switchable input resolutions.

2 Related Work

We briefly summarize related methodologies in previous literature and analyze their relations and differences with our approach as follows.

Scale Deviation. FixRes [30] sheds light on the distribution shift between the train and test data, and quantitatively analyzes its effect on apparent object sizes and activation statistics, which arises from inconsistent data pre-processing

protocols during training and testing time. The discrepancy between train and test resolution is defined as *scale deviation* in this work. We would like to clarify that the issue of scale deviation exists across universally scaled images in the visual classification task, which shows a clear distinction compared to another popular phenomenon named *scale variance*. The problem of *scale variance* is more commonly identified among instances of different sizes within a single image, especially in the images of the MS COCO [19] benchmark for object detection. Typically hierarchical or parallel multi-scale feature fusion approaches [29, 18, 39, 17, 38, 6] are utilized to address this problem, which has a loose connection with our research focus nevertheless. To handle the initial problem of *scale deviation*, FixRes proposes a simple yet effective strategy that prefers increased test crop size and fine-tunes the pre-trained model on the training data with the test resolution as a post-facto compensation. Notably, FixRes lays emphasis on calibrating BN statistics over the training data modulated by the test-size distribution, which is of vital importance to remedy the activation distribution shift. In comparison, we could use a proxy or data-free inference method to avoid the calibration of BN statistics, thus no post-processing steps are involved after end-to-end network training. We further provide empirical comparison in Section 4. Specializing BN layers for network adaption between a few different tasks has been adopted in domain adaption [25], transfer learning [21], adversarial training [31] and optimization of sub-models in a super net [34][33]. Inspired by them, we also overcome the shortcoming of statistic discrepancy through privatizing statistics and trainable parameters in the BN layers of each main network.

Meta Learning. Meta learning, or learning to learn, has come into prominence progressively in the field of artificial intelligence, which is supposed to be the learning mechanism in a more human-like manner. The target of meta learning is to advance the learning procedure at two coupled levels, enabling a *meta learning* algorithm to adapt to unseen tasks efficiently via utilizing another *base learning* algorithm to extract transferable prior knowledge within a set of auxiliary tasks. The hypernetwork [9] is proposed to efficiently generate the weights of the large main network using a small auxiliary network, in a relaxed form of weight-sharing across layers. Bertinetto *et al.* [4] concentrate on the one-shot learning problem and use a *learnnet* to map a single training exemplar to weights of the main network in one go. Andrychowicz *et al.* [1] replace the standard optimization algorithm with an LSTM optimizer to generate an update for the main network in a self-adaptive manner. Meta-LSTM [24] further develops this idea by revealing the resemblance between cell states in an LSTM and network weights in a CNN model with respect to their update process. SMASH [5] applies the HyperNet to Neural Architecture Search (NAS) by transforming a binary encoding of the candidate network architecture to its corresponding weights. In our framework, meta networks, or to say hypernetworks, are responsible for generating weights of convolutional layers in the main network according to scale-related meta knowledge of the input data.

Knowledge Distillation. Knowledge Distillation (KD) is based on a teacher-student knowledge transfer framework in which the optimization of a lower-

capacity student neural network is guided by imitating the soft targets [11][37] or intermediate representation [26][32][35] from a large and powerful pre-trained teacher model. Inspired by curriculum learning [3], RCO [15] promotes the student network to mimic the entire route sequence that the teacher network passed by. RKD [22] and CCKD [23] both exploit structural relation knowledge hidden in the embedding space to achieve correlation congruence. Several recent works also demonstrate the effectiveness of KD in improving the teacher model itself by self-distillation [2][8]. We introduce the scale distillation into our framework to transfer the scale and structural information of the same object instances from large input images to smaller ones at each step throughout the whole training procedure. The knowledge distillation process emerges among different main networks with the same network structure that travel along the same route sequence of optimization, without the assistance of external teacher models.

3 Approach

A schematic overview of our method is presented in Fig. 2. The key innovations lie in the employment of meta networks in the body and scale distillation at the end. In this section, we elaborate on the insights and formulations of them.

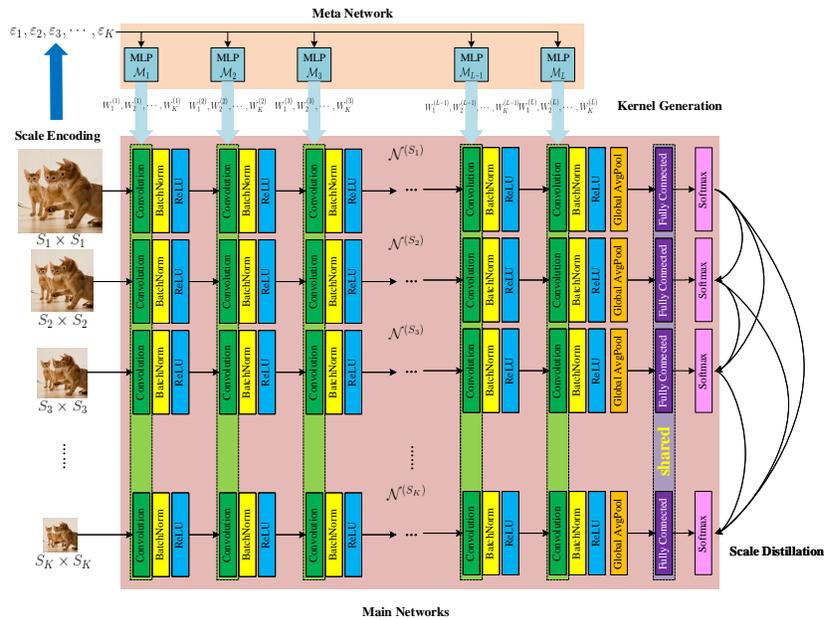


Fig. 2. Schematic illustration of our proposed Scale Adaptive Network framework. The input image scales S_1, S_2, \dots, S_K are linearly transformed into a set of encoding scalars $\epsilon_1, \epsilon_2, \dots, \epsilon_K$, which are fed into the MLP meta network $\mathcal{M}_l, l = 1, 2, \dots, L$, generating weights $\{W_1^{(l)}, W_2^{(l)}, \dots, W_K^{(l)}\}$ of the l^{th} convolutional layer for each main network $\mathcal{N}^{(S_i)}$ associated with one certain input scale $S_i, i = 1, 2, \dots, K$. BN layers are privatized and FC layers are shared among different main networks. The Scale Distillation process is performed in a top-down fashion. Best viewed in electronic version.

3.1 Network Architecture

Motivated by the target of achieving the scale-adaptive peculiarity of CNNs, we speculate that scale-invariant knowledge should be exploited regarding each input image in its heterogeneous modes of scales. Although the same object instances pose significant scale variation in these modes, they could share common structural relations and geometric characteristics. A meta learner built upon the top of CNNs is expected to extract and analyze this complementary knowledge across scales. Then it injects this prior knowledge into underlying CNNs through parameterization, making them quickly adapt to a wide range of new tasks of visual classification which may include unseen scale transformations.

In this spirit, we denote the model optimized on training images with the resolution $S \times S$ as $\mathcal{N}^{(S)}$ and consider a group of main networks optimized on a set of input resolutions $\mathbb{S} = \{S_1, S_2, \dots, S_k\}$ respectively. We also construct a cohort of meta networks $\mathcal{M} = \{\mathcal{M}_1, \mathcal{M}_2, \dots, \mathcal{M}_L\}$ to generate convolutional kernels conditioned on the scale S of input images for an L -layer convolutional neural network $\mathcal{N}^{(S)}$. As illustrated in Fig. 2, bidirectional information flow is established between the shared meta network and each individual main network by the means of **Scale Encoding** and **Kernel Generation** respectively. All meta networks are instantiated by the same network structure of a Multi-Layer Perceptron (MLP).

Scale Encoding. The input to each MLP is an encoding scalar that contains information about the relative scale of training examples. Since the down-sampling rate of prevalent CNNs is $1/32$, we heuristically apply a normalization policy to linearly transform the input scale S to an encoding $\varepsilon = 0.1 \times S/32$.

Kernel Generation. The output dimension of an MLP equals the dimension of its corresponding convolutional kernel but in the flattened form. For example, a 1-in, $(C_{out}C_{in}K^2)$ -out MLP is built to adaptively generate C_{out} groups of convolutional kernels, each group containing C_{in} kernels with the same size of $K \times K$. Compared with hypernetworks [9] that map each learnable embedding vector to the weights of its corresponding layer with one auxiliary network, our mechanism assigns each meta learner to generate weights for the corresponding layer with one common input encoding scalar regarding a specific main network.

With these well-defined meta learners, parameters of convolutional layers in the main network $\mathcal{N}^{(S)}$ can be generated and associated with a specific input scale S . As emphasized by previous works [30], BN layers should be delicately handled for the sake of inconsistency between data distributions of varying scales. Note that parameters in the BN layers usually occupy a negligible portion of the parameters within the whole model, *e.g.*, less than 1% of the total amount in most cases, we opt for a straightforward yet effective strategy by maintaining individual BN layers, denoted as $\text{BN}^{(S)}$, for each scale-specified main network $\mathcal{N}^{(S)}$. Injecting these conditional normalization layers can lend more inherent adaptability and scalability to intermediate feature representations to accommodate scale discriminative knowledge. By contrast, with regard to the last FC layer that occupies a considerable amount of parameters, a shared one is designated to fit any potential main networks.

To take full advantage of various sources of scale information, we propose a mixed-scale optimization scheme accordingly. With the label-preserved transformation, each image in the training set $\mathcal{D} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_N, y_N)\}$ is resized to a series of scales, *e.g.*, S_1, S_2, \dots, S_k , where $y_j, j = 1, 2, \dots, N$ is the ground truth category label of the sampled image \mathbf{x}_j . For a certain scale S_i , we encode it as $\varepsilon^{(S_i)}$ and map the encoding scalar to a fully parameterized main network $\mathcal{N}^{(S_i)}$ through the meta learners. Transformed samples with the same size S_i are assembled to form a resized version of the original training set denoted as $\mathcal{D}^{(S_i)}$ and fed into the corresponding main network $\mathcal{N}^{(S_i)}$. For each pair of $\mathcal{D}^{(S_i)}$ and $\mathcal{N}^{(S_i)}$, it follows the standard optimization procedure of convolutional neural networks via minimizing the cross-entropy loss. Since our objective is to optimize the overall accuracy under different settings of scales, no scale is privileged and the total classification loss is an un-weighted sum of the individual losses, represented as

$$\mathcal{L}_{\text{CE}} = \sum_{i=1}^k \sum_{j=1}^N \mathcal{L}_{\text{CE}}(\boldsymbol{\theta}^{(S_i)}; \mathbf{x}_j^{(S_i)}, y_j), \quad (1)$$

where k is the number of resize transformations and $\boldsymbol{\theta}^{(S_i)} = \{\mathcal{M}_l(\varepsilon^{(S_i)}) \mid l = 1, 2, \dots, L\} \cup \{\text{BN}^{(S_i)}, \text{FC}\}$ are weights of the network $\mathcal{N}^{(S_i)}$ where convolutional layers are directly generated by meta learners. The parameters in meta networks \mathcal{M} are optimized simultaneously following the chain rule as the aforementioned weight generation operations are completely differentiable.

The number of hidden layers and units in the MLP could be tuned. In our main experiments, the meta network is chosen as a single layer MLP for the purpose of effectiveness and efficiency (validated by ablation experiments later). It could be represented as an FC layer with the weight \mathbf{W}_l and bias \mathbf{b}_l

$$\mathcal{M}_l(\varepsilon^{(S_i)}) = \varepsilon^{(S_i)} \mathbf{W}_l + \mathbf{b}_l. \quad (2)$$

Due to the existence of the bias term, the output convolutional kernels are *not* simply distinct up to a scaling factor $\varepsilon^{(S_i)}$ across different main networks. To be cautious, we also examine the value of \mathbf{W}_l and \mathbf{b}_l regarding each layer. They are of the same size and the same order of magnitude in almost all cases, indicating that the weights and biases have an equivalently important influence on the learning dynamics. The ratio of \mathbf{W}_l to \mathbf{b}_l in each layer is reported in the supplementary materials.

3.2 Scale Distillation

High-resolution representations [28] can contain finer local feature descriptions and more discernable semantic meaning than the lower-resolution ones, hence it is appropriate to utilize the probabilistic prediction over larger-scale inputs to offer auxiliary supervision signals for smaller ones, which will be referred to as *Scale Distillation* in this context. Specifically, a Kullback-Leibler divergence $D_{\text{KL}}(\cdot \parallel \cdot)$ is calculated between each pair of output probability distributions

among all main networks, which leads to an additional representation mimicking loss shaped in a top-down manner as follows

$$\mathcal{L}_{\text{SD}} = \sum_{i=1}^k \sum_{\substack{j=1 \\ S_j < S_i}}^k D_{\text{KL}}(\mathbf{p}^{(S_i)} \parallel \mathbf{p}^{(S_j)}), \quad (3)$$

where $\mathbf{p}^{(S_i)}$ denotes the probabilistic distribution prediction with respect to object categories outputted by the main network $\mathcal{N}^{(S_i)}$.

Our mixed-scale training mechanism naturally supports scale distillation as an in-place operation. During each training step, we take the predicted labels of one model and fix them as the soft training labels for other models processing smaller input samples, which can be implemented on the fly without introducing further memory overheads in practice. Compared with conventional KD methodology [11], a main network in our framework may be both the teacher model and the student model, depending on its matched counterpart model. Furthermore, the cohort of main networks are of the same model size and optimized in a single-shot rather than two-stage manner.

Then, the overall optimization objective of our framework is to minimize the combined loss function

$$\mathcal{L} = \alpha \mathcal{L}_{\text{CE}} + \beta \mathcal{L}_{\text{SD}}, \quad (4)$$

where α and β are positive weight coefficients to balance the cross-entropy loss and the scale distillation loss. In our main experiments, we do not make much investment in tuning these hyperparameters and find that simply setting $\alpha = \beta = 1$ leads to satisfactory performance, which demonstrates the robustness of our proposed optimization scheme in some sense.

Intuitively condensing networks for different purposes into a shared framework tends to bring about performance degradation compared to individually trained ones, since they might exhibit inconsistent learning dynamics. However, we surmise that our performance improvements could originate from a relaxed form of knowledge transfer across different scales. According to Eqn. 2, weights and biases of the meta networks respectively enforce convolution parameter scaling and sharing across different main networks. The generated weights for one scale also depend on the information from any other different scales due to the joint training. By feat of the shared meta networks and a collaborative training regime, the knowledge interaction process between models may be interpreted as an *implicit distillation*. In addition to the implicit information sharing mechanism above, we further develop an *explicit distillation* technique to aggressively advance this knowledge transfer process, presented as Scale Distillation.

3.3 Inference

For inference, let the selected training resolution range be $\mathbb{S} = \{S_1, S_2, \dots, S_k\}$ and the test resolution be T , we first search the nearest resolution $S(T) \in \mathbb{S}$ for T , then feed the scale encoding $\varepsilon^{(T)}$ to the pre-trained meta network to parameterize

convolutional layers for the main network and the BN layers reserved for $S(T)$ during training could be applied directly. Finally the *ideal inference* is realized by sending the test image to the network parameterized by

$$\boldsymbol{\theta}_{\text{ideal}}^{(T)} = \{\mathcal{M}_l(\varepsilon^{(T)}) | l = 1, 2, \dots, L\} \cup \{\text{BN}^{(S(T))}, \text{FC}\}. \quad (5)$$

If $S(T)$ is *not* close enough to T regarding its value, the trainable parameters of BN (scale and shift parameter) could still be applied directly but calibration of BN statistics (running mean and variance) is necessary for retaining a decent performance (will be shown later in the ablation studies). Intriguingly, we find that simply substituting the scale encoding $\varepsilon^{(T)}$ by $\varepsilon^{(S(T))}$ to match its corresponding BN layers is ready to make compensation. It says that the runtime network could also be dynamically parameterized by

$$\boldsymbol{\theta}_{\text{proxy}}^{(T)} = \{\mathcal{M}_l(\varepsilon^{(S(T))}) | l = 1, 2, \dots, L\} \cup \{\text{BN}^{(S(T))}, \text{FC}\}, \quad (6)$$

to achieve *approximate* performance as that above (recognition accuracy by using $\boldsymbol{\theta}_{\text{ideal}}^{(T)}$ and $\boldsymbol{\theta}_{\text{proxy}}^{(T)}$ for inference is comprehensively compared in the ablation experiments). Since using $\boldsymbol{\theta}_{\text{proxy}}^{(T)}$ for inference would be relatively convenient once the pre-trained main network weights $\mathcal{M}(\varepsilon^{(S(T))})$ are already on hand, we opt for this *proxy inference* method as an alternative for performance benchmark throughout Section 4, if no further specification. In practice, clients could choose either option to achieve similar performance, depending on whether the desired convolutional weights are handily accessible or even on hand.

Summarily, the **ideal inference** in Eqn. 5 uses test-specified encoding and calibration for BN statistics (if necessary) while the **proxy inference** in Eqn. 6 uses training-specified encoding and no calibration. The concrete algorithms of optimization and inference are provided in the supplementary materials.

4 Experiments

We present extensive experimental results on ImageNet using various prevailing CNN architectures and conduct controlled experiments for introspection.

4.1 Main Results

Our method is evaluated on the large-scale ImageNet [7] dataset, which is a very challenging image recognition benchmark including over 1.2 million training images and 50,000 validation images belonging to 1,000 object categories. We follow the standard practice for data augmentation [10], utilizing random resizing and cropping operations together with horizontal flipping to generate image patches with the desired resolutions. During evaluation, we crop the center region from each transformed image of which the shorter side is resized to satisfy the crop ratio of 0.875. We report top-1 validation error in all the following tables.

Several different choices of main networks are explored to demonstrate the effectiveness and scalability of our approach. Specifically, we select three network

Table 1. Comparison of ResNet-18, MobileNetV2 and ResNet-50 (from top to bottom) baseline models (left panel) and SAN (right panel) on a spectrum of test resolutions.

| Train \ Test | 256 | 224 | 208 | 192 | 176 | 160 | 144 | 128 | 112 | 96 | 64 | Train \ Test | 256 | 224 | 208 | 192 | 176 | 160 | 144 | 128 | 112 | 96 | 64 |
|--------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|--------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| 224 | 71.90 | 70.97 | 69.14 | 68.72 | 66.20 | 64.68 | 60.36 | 56.85 | 50.17 | 42.55 | 20.40 | 224 | 72.96 | 72.65 | 71.89 | 71.89 | 70.39 | 69.62 | 67.15 | 65.36 | 59.96 | 55.57 | 34.98 |
| 192 | 71.71 | 71.25 | 69.64 | 69.75 | 67.70 | 67.04 | 63.23 | 60.47 | 54.21 | 47.73 | 25.10 | 192 | 72.60 | 72.79 | 72.10 | 72.32 | 71.09 | 70.77 | 68.44 | 67.25 | 62.61 | 58.77 | 38.76 |
| 160 | 70.72 | 70.70 | 69.50 | 70.14 | 68.47 | 68.48 | 65.61 | 63.80 | 58.54 | 53.78 | 31.03 | 160 | 71.51 | 72.15 | 71.61 | 72.12 | 71.04 | 71.31 | 69.36 | 68.60 | 64.69 | 61.53 | 43.02 |
| 128 | 67.41 | 68.51 | 67.83 | 69.06 | 67.81 | 68.56 | 66.51 | 66.36 | 61.88 | 58.29 | 38.58 | 128 | 69.04 | 70.30 | 69.78 | 70.90 | 70.19 | 71.01 | 69.35 | 69.28 | 65.96 | 63.87 | 47.33 |
| 96 | 54.68 | 58.87 | 58.46 | 62.39 | 61.62 | 64.84 | 63.42 | 65.51 | 62.39 | 62.56 | 47.48 | 96 | 62.96 | 65.52 | 65.37 | 67.65 | 66.89 | 68.89 | 67.60 | 68.73 | 65.78 | 65.30 | 51.32 |
| Train \ Test | 256 | 224 | 208 | 192 | 176 | 160 | 144 | 128 | 112 | 96 | 64 | Train \ Test | 256 | 224 | 208 | 192 | 176 | 160 | 144 | 128 | 112 | 96 | 64 |
| 224 | 73.04 | 72.19 | 71.13 | 70.09 | 68.48 | 66.10 | 63.13 | 58.63 | 52.50 | 43.91 | 21.23 | 224 | 73.26 | 72.80 | 72.24 | 71.47 | 70.35 | 68.64 | 66.50 | 63.08 | 58.43 | 51.01 | 28.64 |
| 192 | 72.67 | 72.26 | 71.47 | 71.08 | 69.68 | 68.18 | 65.84 | 61.61 | 56.86 | 48.81 | 25.37 | 192 | 72.91 | 72.86 | 72.42 | 72.22 | 71.33 | 70.26 | 68.43 | 65.83 | 61.81 | 55.44 | 33.24 |
| 160 | 70.83 | 71.39 | 70.17 | 71.08 | 69.31 | 69.50 | 66.63 | 65.05 | 59.60 | 54.08 | 30.56 | 160 | 71.68 | 72.16 | 71.79 | 72.31 | 71.63 | 71.16 | 69.67 | 67.96 | 64.61 | 59.62 | 38.75 |
| 128 | 66.70 | 68.18 | 68.19 | 69.14 | 68.65 | 68.98 | 67.78 | 66.74 | 63.73 | 58.69 | 37.14 | 128 | 68.61 | 70.13 | 69.94 | 71.16 | 70.51 | 70.94 | 69.80 | 69.14 | 66.55 | 63.05 | 44.84 |
| 96 | 51.48 | 56.73 | 57.06 | 60.85 | 61.13 | 63.99 | 63.55 | 65.13 | 63.40 | 62.70 | 46.15 | 96 | 60.55 | 64.11 | 64.21 | 66.91 | 66.80 | 68.58 | 67.82 | 68.41 | 66.59 | 65.07 | 50.41 |
| Train \ Test | 256 | 224 | 208 | 192 | 176 | 160 | 144 | 128 | 112 | 96 | 64 | Train \ Test | 256 | 224 | 208 | 192 | 176 | 160 | 144 | 128 | 112 | 96 | 64 |
| 224 | 77.95 | 77.15 | 76.43 | 75.47 | 74.29 | 72.44 | 70.32 | 66.85 | 62.38 | 54.84 | 32.10 | 224 | 78.93 | 78.57 | 78.32 | 77.81 | 77.02 | 76.11 | 74.67 | 72.47 | 69.44 | 64.30 | 44.91 |
| 192 | 77.56 | 77.26 | 76.81 | 76.41 | 75.27 | 74.39 | 72.32 | 69.77 | 65.70 | 59.22 | 37.23 | 192 | 78.63 | 78.67 | 78.43 | 78.44 | 77.64 | 77.10 | 75.86 | 74.00 | 71.29 | 67.05 | 48.91 |
| 160 | 76.69 | 76.93 | 76.60 | 76.61 | 75.81 | 75.31 | 73.91 | 72.12 | 69.11 | 63.87 | 43.74 | 160 | 77.66 | 78.07 | 77.99 | 78.26 | 77.73 | 77.60 | 76.60 | 75.35 | 73.15 | 69.49 | 53.62 |
| 128 | 74.14 | 75.13 | 75.16 | 75.66 | 75.46 | 75.53 | 74.63 | 73.53 | 71.58 | 67.67 | 51.40 | 128 | 75.18 | 76.22 | 76.39 | 77.17 | 77.01 | 77.38 | 76.50 | 76.04 | 74.24 | 71.56 | 57.84 |
| 96 | 68.81 | 70.97 | 71.29 | 72.68 | 72.66 | 73.64 | 73.43 | 73.53 | 72.35 | 70.61 | 59.09 | 96 | 69.91 | 72.24 | 72.69 | 74.18 | 74.33 | 75.39 | 75.12 | 75.30 | 74.28 | 72.83 | 62.14 |

Table 2. Comparison of ResNet-18-based SAN, evaluated on interpolated resolutions with different inference configurations, including (i) ideal inference: before (left) and after (middle) BN calibration; (ii) proxy inference: no need for BN calibration (right).

| Train \ Test | 208 | 176 | 144 | 112 | Train \ Test | 208 | 176 | 144 | 112 | Train \ Test | 208 | 176 | 144 | 112 |
|--------------|-------|-------|-------|-------|--------------|-------|-------|-------|-------|--------------|-------|-------|-------|-------|
| 224 | 71.71 | 67.44 | 57.47 | 39.46 | 224 | 72.03 | 70.32 | 66.90 | 59.18 | 224 | 71.89 | 70.39 | 67.15 | 59.96 |
| 192 | 71.51 | 70.72 | 65.22 | 52.15 | 192 | 72.13 | 71.13 | 68.48 | 62.14 | 192 | 72.10 | 71.09 | 68.44 | 62.61 |
| 160 | 66.21 | 70.52 | 69.07 | 61.10 | 160 | 71.41 | 71.07 | 69.40 | 64.59 | 160 | 71.61 | 71.04 | 69.36 | 64.69 |
| 128 | 49.45 | 64.49 | 68.78 | 65.57 | 128 | 68.95 | 69.89 | 69.52 | 66.16 | 128 | 69.78 | 70.19 | 69.35 | 65.96 |
| 96 | 19.71 | 45.75 | 61.74 | 65.26 | 96 | 62.49 | 65.86 | 67.35 | 66.26 | 96 | 65.37 | 66.89 | 67.60 | 65.78 |

architectures including ResNet [10] with 18/50 layers and MobileNetV2 [27] in view of their strong track record. It is noted that we consider both large-scale networks and a very lightweight one, which also feature regular and inverted residual blocks respectively. Furthermore, we prove that our meta learners could smoothly learn to generate the kernels of both standard convolutions and depth-wise separable convolutions. The ResNet family is trained using the default SGD optimizer with the momentum of 0.9 and the weight decay of 1e-4 for 120 epochs. The learning rate initiates from 0.1 and decays to zero following a half cosine annealing schedule. The batch size is set to 256. The lightweight MobileNetV2 is trained using almost the same optimization hyperparameters and learning rate decay strategy but with a smaller initial learning rate of 0.05 and a smaller weight decay of 4e-5 to suppress underfitting. The optimization procedure lasts for 150 epochs for full convergence. All baselines and our SAN-based models are trained using the above scheme for fair comparisons. The training resolutions of SAN models are set to $\mathbb{S} = \{S_1, S_2, S_3, S_4, S_5\} = \{224, 192, 160, 128, 96\}$.

We take independently trained models as the baselines and evaluate them among a wide range of test resolutions on the ImageNet validation set. The baseline results of three networks are shown in the left panel of Table 1. We report corresponding results using our SAN in the right panel of Table 1, where the results in the j^{th} row is evaluated with the scale encoding $\varepsilon^{(S_j)}$ and BN layers $\text{BN}^{(S_j)}$ without the calibration of statistics. Therefore, the shaded numeric value (in the j^{th} row, s.t. $S(T) = S_j$) points out the inference result using the *proxy inference* method for the test resolution T in each column, as stated in Section 3.3. It is evident that our dynamically parameterized models achieve consistent accuracy improvement over the individually trained baselines. Such

performance enhancement emerges not only on the training resolutions but also on the interpolated and extrapolated ones of the training range, demonstrating the universal applicability of our meta learner. Furthermore, it should be noticed that the generated classifier for one specific resolution also generalizes well on other resolutions (indicated by those numeric values without being shaded in each row) compared to the individually trained model, even obtaining over 10% compensation for those baseline models in several cases with severe scale deviation (around the corner of the table).

The recognition accuracy curves of our SAN with MobileNetV2 and corresponding baseline models are depicted in Fig. 1. It showcases a clear trend that the curves of SAN models envelop those of their baseline counterparts. Our SAN also guarantees much milder performance drop when input samples pose a resolution discrepancy between optimization and inference.

4.2 Ablation Studies

We conduct comprehensive ablation experiments to analyze the influence of different configurations and provide empirical validation for our design.

BN Calibration for Ideal Inference. The encoding scalars used for training are bound to their privatized BN layers. If these scalars are altered according to input resolutions at test time following the ideal inference method, evaluation performance might suffer from incompatibility, as demonstrated in the left panel of Table 2. For compensation, we apply the post-hoc calibration for BN statistics. Specifically, we recalculate the running mean and variance of BN layers over training samples of the test resolution with exact averages rather than moving averages. The evaluation performance after BN calibration is reported in the middle panel of Table 2, showing amelioration for those before calibration. To deserve to be mentioned, when the training resolution approaches the test one, the performance gap would be relatively minor, crediting to the smoothness of the linear meta modeling space and the BN parameter space. Anyway, the above dissection reiterates the critical value of Batch Normalization.

Proxy Inference. We empirically justify that using θ_{proxy} for inference is a credible alternative for model evaluation. For test resolutions included in the training resolution range, *i.e.*, 224, 192, 160, 128 and 96, $S(T) = T$, thus $\theta_{\text{proxy}} = \theta_{\text{ideal}}$ and the inference process will be identical for the two proposed methods. Thus, we lay analytic emphasis on the test resolutions sandwiched between two training resolutions, *i.e.*, 208, 176, 144 and 112, and report their performance in the right panel of Table 2. Since the scale encoding reverts to the ones during training in the proxy inference method, calibration is not needed here. We notice that using proxy inference for each test resolution (as shaded in the right panel of Table 2) leads to nearly the same results as those with the ideal inference method (as shaded in the middle panel of Table 2). These comparisons provide empirical support for our claim in Section 3.3.

Scale Distillation. For the purpose of ablating the influence of our proposed Scale Distillation, we further train a MobileNetV2-based SAN without this technique. Accuracy records of the original SAN are presented in Table 1

Table 3. Accuracy of MobileNetV2-based SAN (w/o SD) (*left*) and MobileNetV2-based SAN (w/o SD) with shared Batch Normalization layers (*right*). Please refer to Table 1 (*middle-right*) for accuracy of intact MobileNetV2-based SAN.

| Train \ Test | 256 | 224 | 208 | 192 | 176 | 160 | 144 | 128 | 112 | 96 | 64 | Train \ Test | 256 | 224 | 208 | 192 | 176 | 160 | 144 | 128 | 112 | 96 | 64 |
|--------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|--------------|-------|-------|-------|-------|-------|-------|-------|-------|--------|-------|-------|
| 224 | 73.14 | 72.53 | 71.50 | 70.94 | 69.26 | 67.91 | 64.50 | 61.84 | 55.18 | 48.60 | 24.94 | 224 | 2.82 | 2.95 | 2.88 | 3.02 | 2.93 | 3.04 | 2.90 | 2.77 | 2.59 | 2.29 | 1.13 |
| 192 | 72.85 | 72.58 | 71.45 | 71.56 | 69.93 | 69.36 | 66.32 | 64.15 | 58.23 | 52.52 | 28.62 | 192 | 27.27 | 29.07 | 29.03 | 29.91 | 29.30 | 29.24 | 28.09 | 25.70 | 23.24 | 18.64 | 8.45 |
| 160 | 71.30 | 71.78 | 70.66 | 71.54 | 69.92 | 70.00 | 67.35 | 66.09 | 60.85 | 56.20 | 32.94 | 160 | 66.41 | 66.89 | 66.46 | 66.52 | 65.62 | 65.02 | 63.08 | 60.61 | 56.812 | 50.66 | 29.64 |
| 128 | 67.42 | 69.11 | 67.64 | 69.95 | 68.37 | 69.65 | 67.43 | 67.19 | 62.64 | 59.54 | 38.26 | 128 | 49.17 | 50.86 | 50.26 | 51.43 | 50.10 | 50.70 | 47.92 | 46.73 | 41.52 | 37.71 | 19.82 |
| 96 | 56.25 | 60.34 | 58.50 | 63.81 | 61.86 | 66.25 | 63.87 | 66.23 | 62.25 | 61.86 | 44.05 | 96 | 0.76 | 0.90 | 0.85 | 1.03 | 0.94 | 1.23 | 1.13 | 1.28 | 1.07 | 1.38 | 1.07 |

Table 4. Accuracy of parallel ResNet-18 models with privatized convolution for each resolution and scale distillation across resolutions. Please refer to Table 1 (*top-left* and *top-right*) for results of the baseline and SAN.

| Train \ Test | 256 | 224 | 208 | 192 | 176 | 160 | 144 | 128 | 112 | 96 | 64 |
|--------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| 224 | 72.06 | 71.10 | 69.54 | 68.87 | 66.23 | 64.54 | 60.47 | 56.89 | 50.11 | 42.60 | 20.32 |
| 192 | 72.53 | 71.94 | 70.41 | 70.79 | 68.28 | 67.93 | 64.17 | 61.97 | 55.33 | 49.39 | 25.96 |
| 160 | 70.93 | 71.37 | 70.34 | 71.21 | 69.55 | 69.82 | 66.94 | 65.62 | 60.57 | 55.92 | 33.51 |
| 128 | 67.57 | 69.11 | 68.56 | 69.89 | 68.78 | 69.89 | 67.81 | 67.87 | 63.84 | 61.46 | 41.96 |
| 96 | 56.80 | 60.81 | 60.75 | 64.31 | 63.53 | 66.67 | 65.22 | 67.32 | 64.35 | 64.53 | 50.62 |

(middle-right) and the accuracy of SAN (w/o SD) is provided in Table 3 for comparison. We observe that the accuracy drop increases in the test scenarios with smaller input scales, which empirically validates the importance of transferring knowledge from high resolutions to lower ones by introducing the extra supervision signal. As a side benefit, the performance of models over interpolated resolutions is ameliorated to a large extent after the application of Scale Distillation, which may be attributed to the improved interaction of multi-scale information among the same image instances with different resolutions.

Privatized BN. Sharing similar philosophy with [30] and [34], we privatize BN layers by design to eliminate incompatibility between various scaled distributions. For the purpose of ablation, we take the above MobileNetV2-based SAN (w/o SD) as an exemplar and substitute shared BN layers for privatized ones. The consequent accuracy is provided in Table 3, where the benefit of BN is skewed to resolutions in the middle of the training resolution range while the performance in each case still deteriorates compared to its counterpart with privatized BNs. We observe that the training process is stable but the validation accuracy is extremely low because during training the mean and variance of the current mini-batch with a specific resolution are applied while the validation process depends on the moving average statistics from all resolutions. The experimental results further prove the point that a unified BN is insufficient to strike a balance among a broad range of resolutions concurrently.

Parameterized Convolution. A quick question is that why not simply use privatized convolutional layers for each main network rather than employing an auxiliary meta learner? We take the ResNet18 model as an example and show the accuracy with privatized convolutional layers in Table 4 (scale distillation is still applicable in this context). In comparison, the performance are mostly superior to baseline but inferior to our proposed SAN. The meta learning algorithm is adept at integrating prior experience from multiple existing tasks and achieving fast adaption to unseen tasks. In accordance to this rationale, the comparison results also speak to the advantage of utilizing meta networks to collect multi-scale information and dynamically parameterize the convolutional layers for arbitrary image resolutions at runtime.

Table 5. Absolute accuracy drop of ResNet-18-based SAN with non-linear meta learners using 8/16 (*left/right*) hidden units in the FC layer. Please refer to Table 1 (*top-right*) for results of the original SAN equipped with linear meta learners. Note that negative numbers indicate accuracy increment.

| Train \ Test | 256 | 224 | 208 | 192 | 176 | 160 | 144 | 128 | 112 | 96 | 64 | Train \ Test | 256 | 224 | 208 | 192 | 176 | 160 | 144 | 128 | 112 | 96 | 64 |
|--------------|-------|-------|------|-------|------|------|------|-------|------|------|-------|--------------|-------|-------|-------|-------|-------|-------|-------|------|-------|------|------|
| 224 | -0.22 | -0.33 | 0.11 | 0.12 | 0.15 | 0.01 | 0.26 | 0.19 | 0.29 | 0.76 | 1.50 | 224 | -0.48 | -0.31 | -0.24 | 0.01 | -0.17 | -0.11 | 0.21 | 0.67 | 0.55 | 1.35 | 2.20 |
| 192 | -0.21 | -0.07 | 0.19 | 0.25 | 0.27 | 0.20 | 0.36 | 0.56 | 0.92 | 1.30 | 2.27 | 192 | -0.39 | -0.13 | -0.03 | 0.09 | -0.13 | 0.08 | 0.34 | 1.03 | 1.04 | 1.80 | 2.74 |
| 160 | -0.08 | -0.03 | 0.18 | 0.02 | 0.19 | 0.14 | 0.27 | 0.29 | 0.62 | 0.88 | 2.16 | 160 | -0.33 | -0.18 | -0.05 | -0.11 | -0.23 | 0.00 | 0.02 | 0.42 | 0.57 | 1.25 | 2.27 |
| 128 | 0.22 | 0.19 | 0.16 | -0.02 | 0.43 | 0.09 | 0.22 | -0.00 | 0.25 | 0.46 | 1.24 | 128 | 0.11 | 0.11 | -0.11 | -0.06 | -0.12 | 0.08 | -0.26 | 0.03 | 0.03 | 0.42 | 1.32 |
| 96 | 0.93 | 0.80 | 1.00 | 0.54 | 0.76 | 0.36 | 0.66 | 0.26 | 0.09 | 0.11 | -0.07 | 96 | 1.26 | 1.04 | 0.87 | 0.74 | 0.27 | 0.38 | 0.02 | 0.23 | -0.28 | 0.12 | 0.05 |

Architecture of Meta Learner. We explore different MLP structures of the meta network design. Besides the single FC layer adopted in our main experiments, we draw into the non-linear activation to construct a two-layer MLP, where consecutive linear FC layers are interleaved with a non-linear ReLU activation operation and the number of hidden units is set to 8 or 16 in order to avoid introducing heavy computational overheads. The ResNet-18 model is selected for this experiment in view of its relatively low memory consumption. The absolute accuracy drop after substituting the non-linear meta learner with 8 hidden units for the default linear one is summarized in Table 5 (left). We could find that the final performance of these two designs is in close proximity to each other on the medium resolutions, which implies that though conceptually simple, a linear modeling space of meta networks is sufficient for giving rise to desired scale-adaptive characteristics and satisfactory performance. Nevertheless, the decline is not negligible on quite large and small resolutions, which might be attributed to over-fitting problems caused by expansion of the non-linear modeling space. It is probably difficult for the SAN (w/ non-linearity) to generalize well to these situations with large scale deviation. We also explore larger hidden-unit space via doubling its unit number and observe that the tendency of overfitting to marginal resolutions becomes more obvious, as shown in Table 5 (right).

Generalization to Large Resolutions. We shift the range of training resolutions to a larger magnitude, using ResNet-18 as the test case. As shown in Table 6, the accuracy improvement of SAN presents a similar trend within a range of large resolutions $\mathbb{S} = \{384, 320, 256, 224, 192\}$ compared to the smaller ones discussed in Section 4.1. The results empirically show the effectiveness and robustness of our method in handling images with different ranges of resolutions.

Comparison to FixRes. FixRes [30] is merely akin to us in the motivation of alleviating *scale deviation*, but it focuses on further pushing the performance upper boundary of large-scale models with resolution adaption. In stark contrast, we pay attention to enabling efficient and flexible inference conditioned on the input image resolution for evaluation at runtime. A performance comparison is applicable for these two methods using the ResNet-50 model trained on the image resolution of 224×224 , as summarized in Table 7. We are able to beat FixRes when test resolutions are relatively small, which is especially beneficial for resource-limited environments. Furthermore, due to the limitation of available computational resources now, our training resolution range could be at most set to $\{224, 192, 160, 128, 96\}$. If we could enlarge this range, *e.g.*, including training images with larger resolutions such as 384, 448 and 480, our SAN may achieve better performance on these large scales.

Table 6. Comparison of ResNet-18 baseline models (*top*) and ResNet-18-based SAN with a range of larger training resolutions (*bottom*). The shaded numeric values have the same meaning as Table 1.

| Train \ Test | 384 | 320 | 256 | 224 | 208 | 192 | 176 | 160 | 144 | 128 | 112 | 96 | 64 |
|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
| 384 | 73.20 | 71.23 | 67.33 | 63.83 | 60.65 | 58.62 | 54.71 | 50.58 | 45.09 | 38.86 | 32.05 | 23.96 | 8.89 |
| 320 | 73.42 | 72.56 | 70.03 | 67.44 | 64.92 | 63.23 | 59.47 | 56.07 | 51.17 | 45.89 | 38.64 | 30.27 | 11.74 |
| 256 | 72.75 | 72.68 | 71.71 | 70.24 | 68.00 | 67.27 | 64.05 | 62.12 | 57.20 | 53.27 | 45.45 | 38.00 | 17.00 |
| 224 | 71.87 | 72.36 | 71.90 | 70.97 | 69.14 | 68.72 | 66.20 | 64.68 | 60.36 | 56.85 | 50.17 | 42.55 | 20.40 |
| 192 | 69.95 | 71.07 | 71.71 | 71.25 | 69.64 | 69.75 | 67.69 | 67.04 | 63.23 | 60.47 | 54.21 | 47.73 | 25.10 |

| Train \ Test | 384 | 320 | 256 | 224 | 208 | 192 | 176 | 160 | 144 | 128 | 112 | 96 | 64 |
|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
| 384 | 74.69 | 74.12 | 72.14 | 70.28 | 68.65 | 67.06 | 64.77 | 61.97 | 58.25 | 53.41 | 47.33 | 39.51 | 18.66 |
| 320 | 74.61 | 74.50 | 73.28 | 71.72 | 70.39 | 69.07 | 67.10 | 64.64 | 62.28 | 56.94 | 51.22 | 43.63 | 21.85 |
| 256 | 73.41 | 74.05 | 73.71 | 72.87 | 71.85 | 70.86 | 69.43 | 67.53 | 64.42 | 60.86 | 55.56 | 48.26 | 25.80 |
| 224 | 72.27 | 73.39 | 73.59 | 73.00 | 72.29 | 71.52 | 70.16 | 68.52 | 65.64 | 62.32 | 57.36 | 50.54 | 28.13 |
| 192 | 70.44 | 72.15 | 73.11 | 72.73 | 72.21 | 71.74 | 70.72 | 69.36 | 66.74 | 63.80 | 59.20 | 52.72 | 30.52 |

Table 7. Comparison of FixRes and our SAN using ResNet-50 with training image size of 224×224 on ImageNet. The results of FixRes are extracted from the original publication. The better result for each test resolution is highlighted in **bold**.

| Method \ Test resolution | 64 | 128 | 224 | 288 | 384 | 448 | 480 |
|--------------------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| FixRes [30] | 41.7 | 67.7 | 77.1 | 78.6 | 79.0 | 78.4 | 78.1 |
| SAN (ours) | 44.9 | 72.5 | 78.4 | 78.6 | 77.0 | 75.3 | 74.3 |

More Ablation and Analysis. We provide a great many additional results in the supplementary materials, including analysis of complementary information across resolutions, the superiority of switching input resolutions over network widths, border and round-off effects, results of inference on more dense sampled resolutions, visualization of privatized BN layers, a data-free form of ideal inference, comparison to stronger baselines, and so on.

5 Conclusion

In this paper, we have proposed a Scale Adaptive Network (SAN) framework. For each main network encompassed into the meta learning algorithm, convolutional kernels are dynamically generated by meta networks and BN layers are specially treated. The meta learner could parameterize neural networks for visual classification conditioned on the input resolution at runtime, achieving considerably better performance compared with individually trained models. It is compatible with any CNN backbone architectures, providing an adaptive resolution-accuracy trade-off for fast adaption to environments with different real-time demands. The generality of the proposed framework makes it promising to be translated well to other application domains, such as object detection and semantic segmentation, which is expected to appear in the future work.

Acknowledgement

We sincerely thank Ming Lu and Aojun Zhou from Intel Labs China for their insightful discussions, warm encouragement and valuable feedback. We would like to thank SenseTime Group Limited for their support of computational resource.

References

1. Andrychowicz, M., Denil, M., Gómez, S., Hoffman, M.W., Pfau, D., Schaul, T., Shillingford, B., de Freitas, N.: Learning to learn by gradient descent by gradient descent. In: NIPS (2016)
2. Bagherinezhad, H., Horton, M., Rastegari, M., Farhadi, A.: Label Refinery: Improving ImageNet Classification through Label Progression. arXiv e-prints arXiv:1805.02641 (May 2018)
3. Bengio, Y., Louradour, J., Collobert, R., Weston, J.: Curriculum learning. In: ICML (2009)
4. Bertinetto, L., Henriques, J.a.F., Valmadre, J., Torr, P., Vedaldi, A.: Learning feed-forward one-shot learners. In: NIPS (2016)
5. Brock, A., Lim, T., Ritchie, J., Weston, N.: SMASH: One-shot model architecture search through hypernetworks. In: ICLR (2018)
6. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: ECCV (2018)
7. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A Large-Scale Hierarchical Image Database. In: CVPR (2009)
8. Furlanello, T., Lipton, Z., Tschannen, M., Itti, L., Anandkumar, A.: Born again neural networks. In: ICML (2018)
9. Ha, D., Dai, A.M., Le, Q.V.: Hypernetworks. In: ICLR (2017)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
11. Hinton, G., Vinyals, O., Dean, J.: Distilling the Knowledge in a Neural Network. arXiv e-prints arXiv:1503.02531 (Mar 2015)
12. Howard, A., Sandler, M., Chu, G., Chen, L.C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., Le, Q.V., Adam, H.: Searching for mobilenetv3. In: ICCV (2019)
13. Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H.: MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. arXiv e-prints arXiv:1704.04861 (Apr 2017)
14. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: ICML (2015)
15. Jin, X., Peng, B., Wu, Y., Liu, Y., Liu, J., Liang, D., Yan, J., Hu, X.: Knowledge distillation via route constrained optimization. In: ICCV (2019)
16. Li, D., Zhou, A., Yao, A.: Hbonet: Harmonious bottleneck on two orthogonal dimensions. In: ICCV (2019)
17. Li, Y., Chen, Y., Wang, N., Zhang, Z.: Scale-aware trident networks for object detection. In: ICCV (2019)
18. Lin, T.Y., Dollar, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: CVPR (2017)
19. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common objects in context. In: ECCV (2014)
20. Ma, N., Zhang, X., Zheng, H.T., Sun, J.: Shufflenet v2: Practical guidelines for efficient cnn architecture design. In: ECCV (2018)
21. Mudrakarta, P.K., Sandler, M., Zhmoginov, A., Howard, A.: K for the price of 1: Parameter efficient multi-task and transfer learning. In: ICLR (2019)
22. Park, W., Kim, D., Lu, Y., Cho, M.: Relational knowledge distillation. In: CVPR (2019)

23. Peng, B., Jin, X., Liu, J., Zhou, S., Wu, Y., Liu, Y., Li, D., Zhang, Z.: Correlation congruence for knowledge distillation. In: ICCV (2019)
24. Ravi, S., Larochelle, H.: Optimization as a model for few-shot learning. In: ICLR (2017)
25. Rebuffi, S.A., Bilen, H., Vedaldi, A.: Learning multiple visual domains with residual adapters. In: NIPS (2017)
26. Romero, A., Ballas, N., Kahou, S.E., Chassang, A., Gatta, C., Bengio, Y.: Fitnets: Hints for thin deep nets. In: ICLR (2015)
27. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: Mobilenetv2: Inverted residuals and linear bottlenecks. In: CVPR (2018)
28. Sun, K., Xiao, B., Liu, D., Wang, J.: Deep high-resolution representation learning for human pose estimation. In: CVPR (2019)
29. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: CVPR (2016)
30. Touvron, H., Vedaldi, A., Douze, M., Jégou, H.: Fixing the train-test resolution discrepancy. In: NeurIPS (2019)
31. Xie, C., Tan, M., Gong, B., Wang, J., Yuille, A.L., Le, Q.V.: Adversarial examples improve image recognition. In: CVPR (2020)
32. Yim, J., Joo, D., Bae, J., Kim, J.: A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In: CVPR (2017)
33. Yu, J., Huang, T.S.: Universally slimmable networks and improved training techniques. In: ICCV (2019)
34. Yu, J., Yang, L., Xu, N., Yang, J., Huang, T.: Slimmable neural networks. In: ICLR (2019)
35. Zagoruyko, S., Komodakis, N.: Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In: ICLR (2017)
36. Zhang, X., Zhou, X., Lin, M., Sun, J.: Shufflenet: An extremely efficient convolutional neural network for mobile devices. In: CVPR (2018)
37. Zhang, Y., Xiang, T., Hospedales, T.M., Lu, H.: Deep mutual learning. In: CVPR (2018)
38. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: CVPR (2017)
39. Zhou, P., Ni, B., Geng, C., Hu, J., Xu, Y.: Scale-transferrable object detection. In: CVPR (2018)