ECCV'20 ONLINE
23-28 AUGUST 2020

Resolution Switchable Networks for Runtime Efficient Image Recognition

Yikai Wang[1], Fuchun Sun[1], Duo Li[2], Anbang Yao[2]

[1] Tsinghua University [2] Intel Labs China

RS-Nets: Resolution Switchable Networks for Runtime Efficient Image Recognition

- **Objective:**

  Obtain a single model which can handle different image resolutions during inference.

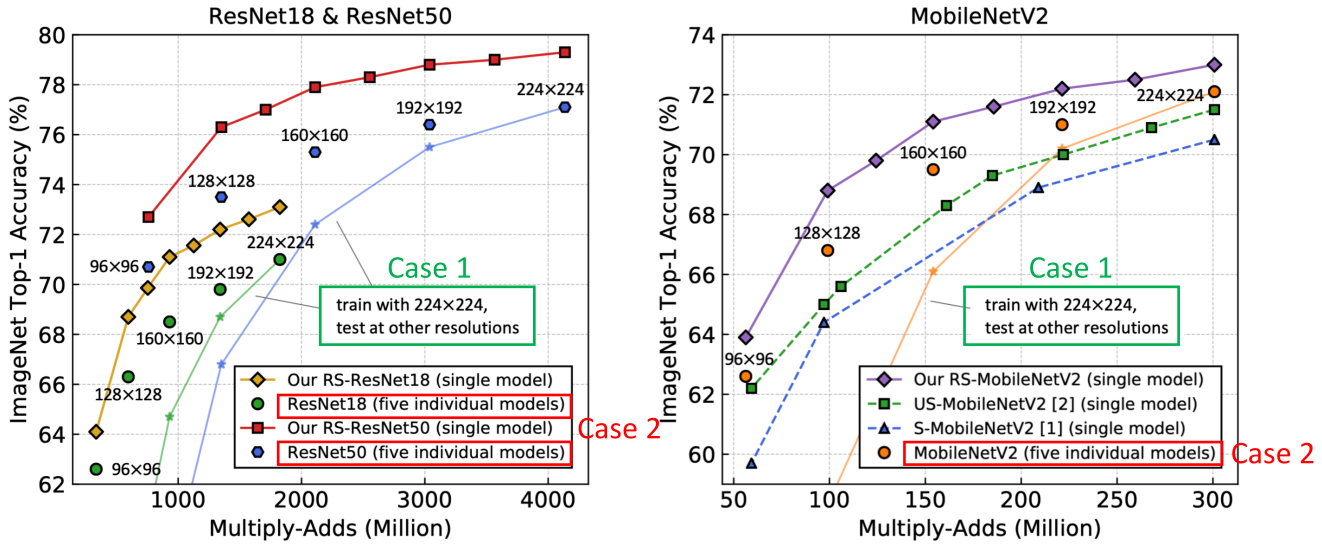
- **Motivation:**

  By switching resolutions, the **running speeds and costs are adjustable** to flexibly handle the real-time latency and power requirements **for different application scenarios or workloads**.

- **Why this is hard for common cases:**

  Case 1: Training a model with a fixed image resolution input——Acc. drops when tested at other resolutions.

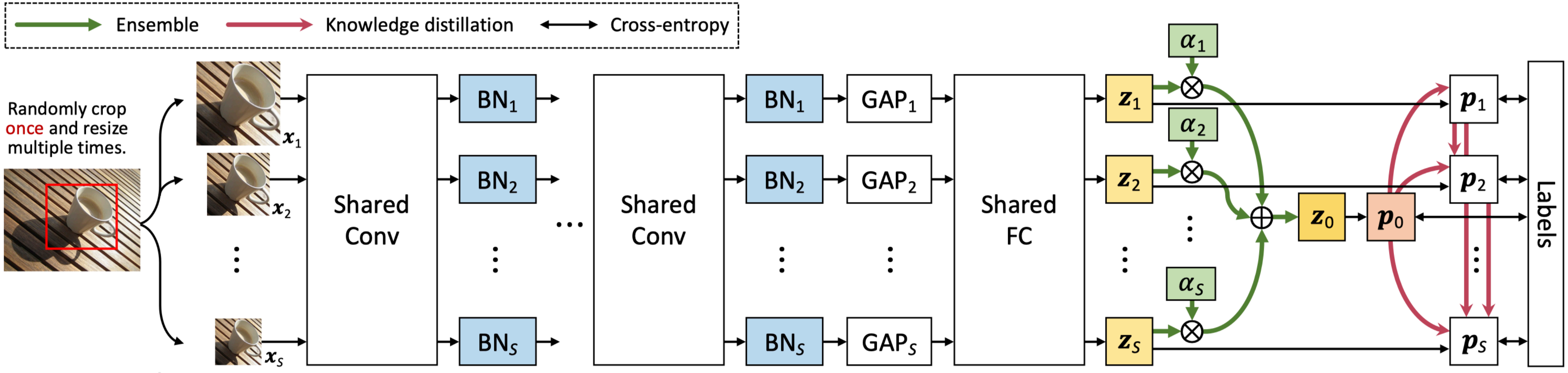
  Case 2: Training an individual model for each image resolution.



[1] Slimmable neural networks. ICLR2019. [2] Universally slimmable networks and improved training techniques. ICCV2019.

RS-Nets: Resolution Switchable Networks for Runtime Efficient Image Recognition

- **Key elements of the paper:**



1. Basic framework——share parameters yet privatize BNs.
2. Analysis of multi-resolution interaction effects.
3. On-the-fly ensemble and knowledge distillation.

RS-Nets: Resolution Switchable Networks for Runtime Efficient Image Recognition

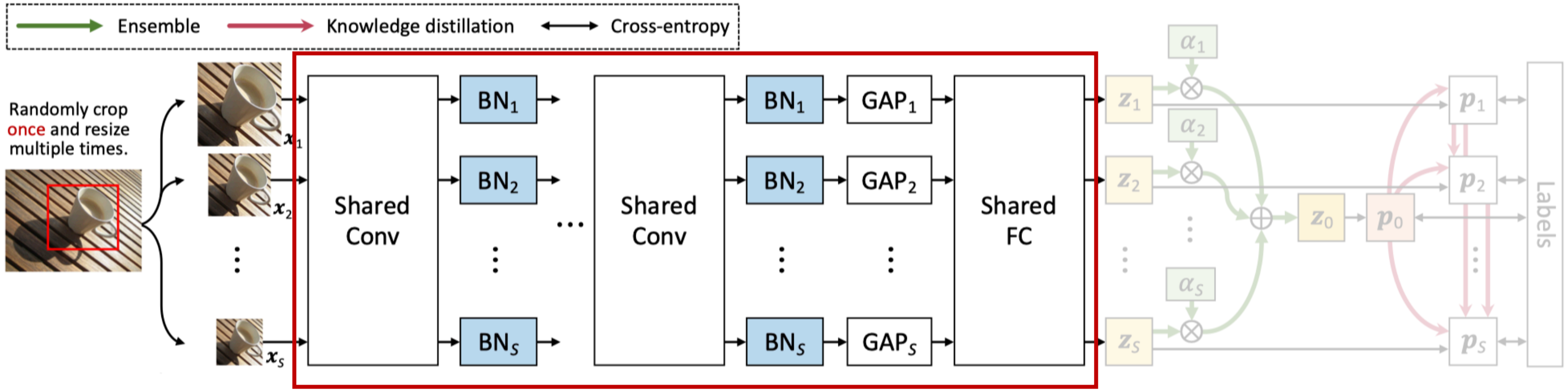
1. Basic framework:

For multi-resolution training, the total loss function is the sum of the cross-entropy losses,

$$\mathcal{L}_{cls} = \sum_{s=1}^{S} \mathcal{H}(\boldsymbol{x}_s, \boldsymbol{y}) \,, \qquad \text{where} \quad \mathcal{H}(\boldsymbol{x}, \boldsymbol{y}) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} \delta(c, y^i) \log\left(p(c|\boldsymbol{x}^i, \boldsymbol{\theta})\right)$$

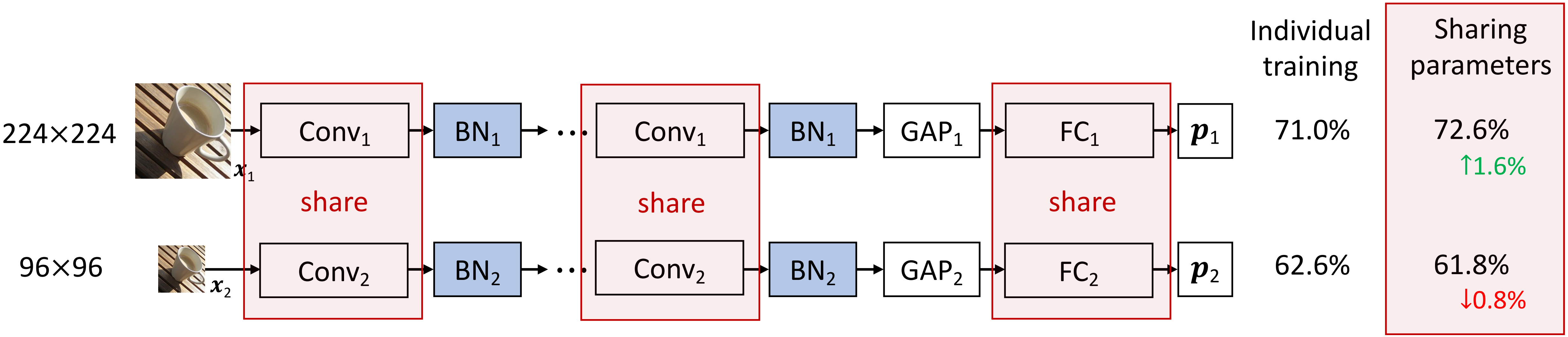
Share parameters yet privatize Batch Normalization layers, for the sth resolution:

$$\boldsymbol{y}_s' = \gamma_s \frac{\boldsymbol{y}_s - \boldsymbol{\mu}_s}{\sqrt{\boldsymbol{\sigma}_s^2 + \epsilon}} + \boldsymbol{\beta}_s, s \in \{1, 2, \cdots, S\}$$



RS-Nets: Resolution Switchable Networks for Runtime Efficient Image Recognition

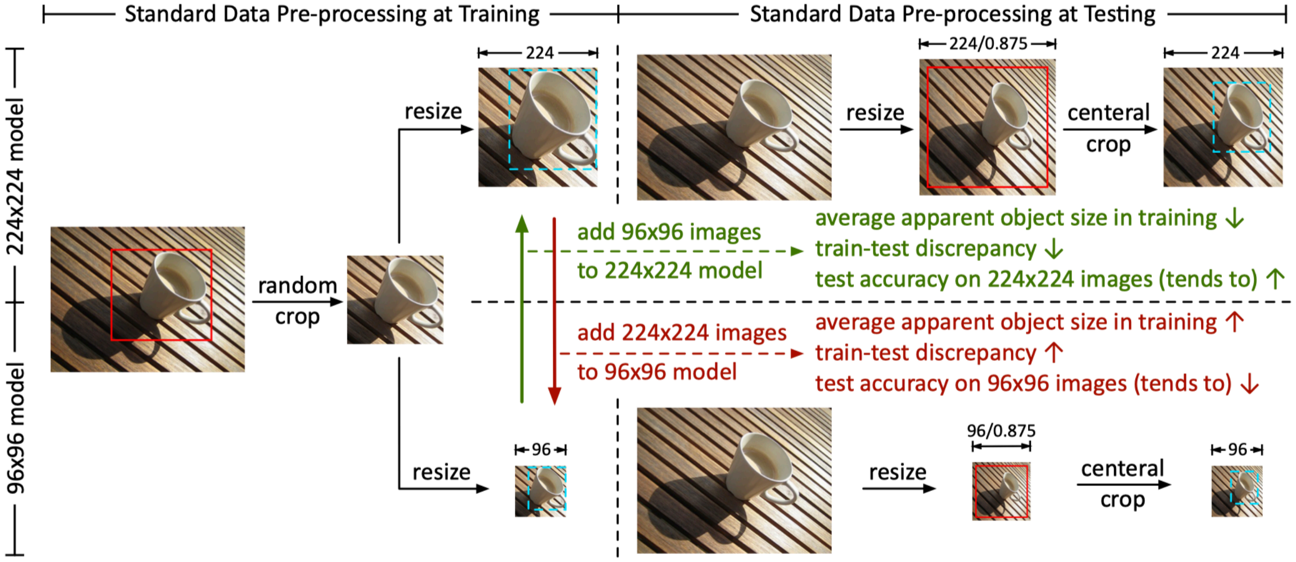
- **2. Analysis of multi-resolution interaction effects**
  - For image recognition, usually a large image resolution corresponds to a high accuracy.
  - E.g, for ResNet18, we get 71.0% and 62.6% top-1 accuracies for 224×224 and 96×96 respectively.
  - What if we share the model parameters, including all Conv layers and the FC layer?
  - The result is, by simply adding the 96×96 resolution images for co-training, accuracy at the 224×224 resolution is obviously improved (+1.6%). But the accuracy at the 96×96 resolution drops (-0.8%).

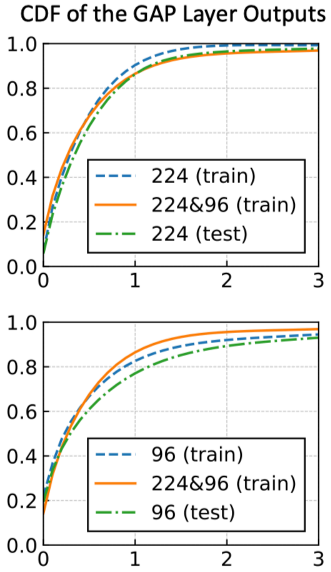


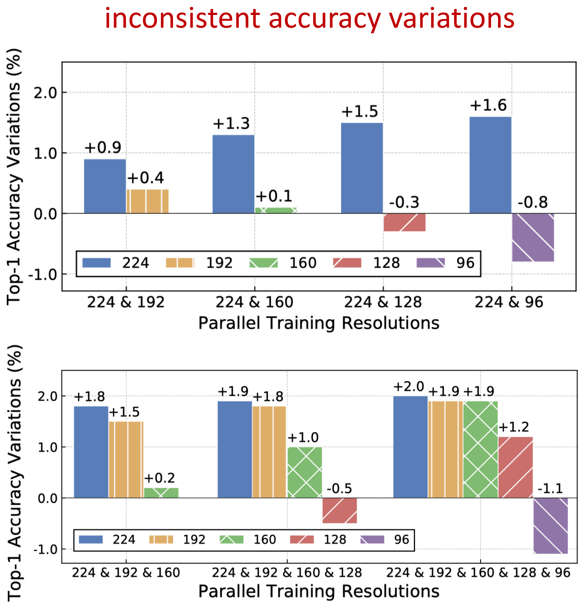
RS-Nets: Resolution Switchable Networks for Runtime Efficient Image Recognition

2. Analysis of multi-resolution interaction effects

- Accuracy gaps over different resolutions tend to be enlarged (not just for the 224×224 and 96×96 case), for which in our paper, we provide an analysis from the aspect of the train-test recognition discrepancy.
- To alleviate such inconsistent accuracy variations, we propose an ensemble distillation design (next page).







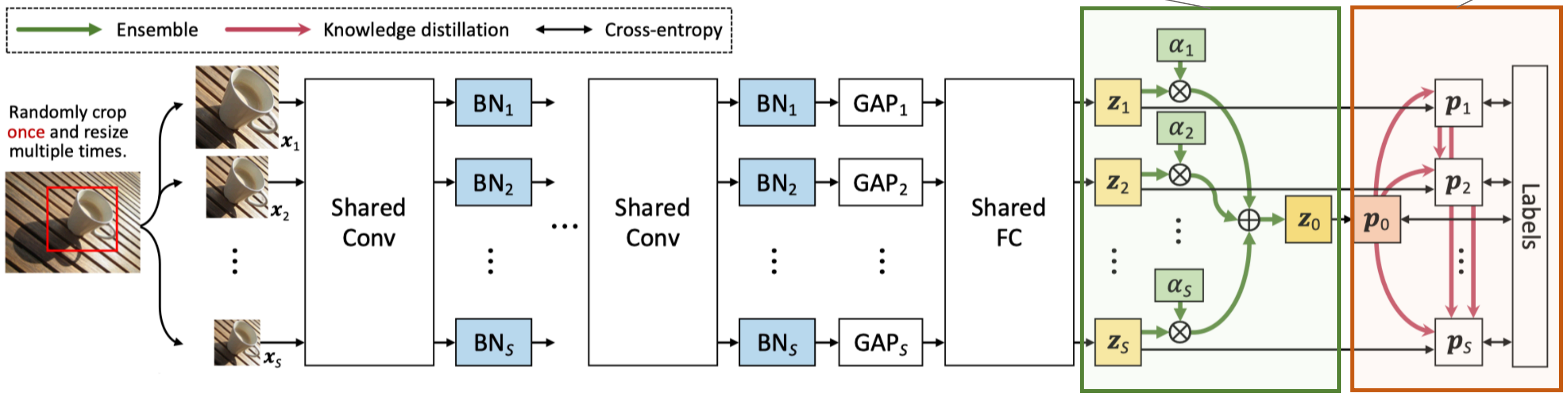
RS-Nets: Resolution Switchable Networks for Runtime Efficient Image Recognition

3. On-the-fly ensemble and knowledge distillation

- A new design of ensemble and knowledge distillation, which can be learnt on-the-fly based on the same image instances with different resolutions. The ensemble and distillation are not needed during inference.

$$z_0 = \sum_{s=1}^{S} \alpha_s z_s \qquad \mathcal{L}_{ens} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C}\delta(c, y^i)\log\left(p(c|z_0^i)\right)$$

$$\mathcal{L}_{dis} = \frac{2}{S+1}\sum_{t=0}^{S-1}\sum_{s=t+1}^{S}\mathcal{D}_{kl}(p_t \| p_s)$$



Total loss:

$$\mathcal{L} = \mathcal{L}_{cls} + \mathcal{L}_{ens} + \mathcal{L}_{dis}$$

ECCV'20 ONLINE 23-28 AUGUST 2020

RS-Nets: Resolution Switchable Networks for Runtime Efficient Image Recognition

- **Experiments and codes**
  - Our code and models are available at https://github.com/yikaiw/RS-Nets
  - We perform experiments on ImageNet (ILSVRC12) dataset. We provide PyTorch implementation for non-quantization and TensorFlow implementation for quantization.

Basic results (non-quantization)

Network	Resolution	MAdds	I-Nets (base)	I-224	Our Parallel	Our RS-Net
ResNet18	224 × 224	1.82G	71.0 / 90.0	71.0 / 90.0	73.0 / 90.9 (+2.0)	**73.1 / 91.0** (+2.1)
	192 × 192	1.34G	69.8 / 89.4	68.7 / 88.5 (-1.1)	71.7 / 90.3 (+1.9)	**72.2 / 90.6** (+2.4)
	160 × 160	931M	68.5 / 88.2	64.7 / 85.9 (-5.2)	70.4 / 89.6 (+1.9)	**71.1 / 90.1** (+2.6)
	128 × 128	596M	66.3 / 86.8	56.8 / 80.0 (-9.5)	67.5 / 87.8 (+1.2)	**68.7 / 88.5** (+2.4)
	96 × 96	335M	62.6 / 84.1	42.5 / 67.9 (-20.1)	61.5 / 83.5 (-1.1)	**64.1 / 85.3** (+1.5)
	Total Params		55.74M	11.15M	11.18M	11.18M
ResNet50	224 × 224	4.14G	77.1 / 93.4	77.1 / 93.4	78.9 / 94.4 (+1.8)	**79.3 / 94.6** (+2.2)
	192 × 192	3.04G	76.4 / 93.2	75.5 / 92.5 (-0.9)	78.1 / 94.0 (+1.7)	**78.8 / 94.4** (+2.4)
	160 × 160	2.11G	75.3 / 92.4	72.4 / 90.7 (-2.9)	76.9 / 93.1 (+1.6)	**77.9 / 93.9** (+2.6)
	128 × 128	1.35G	73.5 / 91.4	66.8 / 87.0 (-6.7)	74.9 / 92.1 (+1.4)	**76.3 / 93.0** (+2.8)
	96 × 96	760M	70.7 / 89.8	54.9 / 78.2 (-15.8)	70.2 / 89.4 (-0.5)	**72.7 / 91.0** (+2.0)
	Total Params		121.87M	24.37M	24.58M	24.58M
M-NetV2	224 × 224	301M	72.1 / 90.5	72.1 / 90.5	72.8 / 90.9 (+0.7)	**73.0 / 90.8** (+0.9)
	192 × 192	221M	71.0 / 89.8	70.2 / 89.1 (-0.9)	71.7 / 90.2 (+0.7)	**72.2 / 90.5** (+1.2)
	160 × 160	154M	69.5 / 88.9	66.1 / 86.3 (-3.2)	70.1 / 89.2 (+0.6)	**71.1 / 90.2** (+1.6)
	128 × 128	99M	66.8 / 87.0	58.3 / 81.2 (-8.5)	67.3 / 87.2 (+0.5)	**68.8 / 88.2** (+2.0)
	96 × 96	56M	62.6 / 84.0	43.9 / 69.1 (-18.7)	61.4 / 83.3 (-1.2)	**63.9 / 84.9** (+1.3)
	Total Params		16.71M	3.34M	3.47M	3.47M

Quantization results

Network	Resolution	Bit-width (W/A): 2 / 32		Bit-width (W/A): 2 / 2	
		I-Nets (base)	Our RS-Net	I-Nets (base)	Our RS-Net
Quantized ResNet18	224 × 224	68.0 / 88.0	**68.8 / 88.4** (+0.8)	64.9 / 86.0	**65.8 / 86.4** (+0.9)
	192 × 192	66.4 / 86.9	**67.6 / 87.8** (+1.2)	63.1 / 84.7	**64.8 / 85.8** (+1.7)
	160 × 160	64.5 / 85.5	**66.0 / 86.5** (+1.5)	61.1 / 83.3	**62.9 / 84.2** (+1.8)
	128 × 128	61.5 / 83.4	**63.1 / 84.5** (+1.6)	58.1 / 80.8	**59.3 / 81.9** (+1.2)
	96 × 96	56.3 / 79.4	**56.6 / 79.9** (+0.3)	52.3 / 76.4	**52.5 / 76.7** (+0.2)
Quantized ResNet50	224 × 224	74.6 / 92.2	**76.0 / 92.8** (+1.4)	72.2 / 90.8	**74.0 / 91.5** (+1.8)
	192 × 192	73.5 / 91.3	**75.1 / 92.4** (+1.6)	70.9 / 89.8	**73.1 / 91.0** (+2.2)
	160 × 160	71.9 / 90.4	**73.8 / 91.6** (+1.9)	69.0 / 88.5	**71.4 / 90.0** (+2.4)
	128 × 128	69.6 / 88.9	**71.7 / 90.2** (+2.1)	66.6 / 86.9	**68.9 / 88.3** (+2.3)
	96 × 96	65.5 / 86.0	**67.3 / 87.4** (+1.8)	61.7 / 83.4	**63.4 / 84.7** (+1.7)