Supplementary Materials for: Resolution Adaptive Networks for Efficient Inference

1. Appendix A: Implementation Details

In this section, we introduce the architecture configurations for our RANets and MSDNets in the experiments of the main paper.

1.1. CIFAR-10 and CIFAR-100

MSDNet: For CIFAR-10 and CIFAR-100, features with 3 different scales (32×32, 16×16, 8×8) are used for MSDNets in our experiments. The trained MSDNets have $\{6, 8, 10\}$ classifiers, where their depths are $\{16, 20, 24\}$, respectively.

RANet: The same 3 scales features are utilized for our RANets in the experiments. However, as mentioned in section 3.3.1, different from MSDNet, the scales of the generated base features can be different, and we could have a RANet with three or four base features in three scales. We test 3 architecture configurations as follows:

Model-C-1: The size of three base features are $32 \times 32, 16 \times 16, 8 \times 8$. Three sub-networks corresponding to these base features have $6, 4, 2$ *Conv Blocks*, respectively. We set two step mode for RANet to control the number of layers in each *Conv Block*: 1)even: the number of layers in each *Conv Block* is set to 4; 2)linear growth (lg): the number of layers in a *Conv Block* is added 2 to the previous one, and the base number of layers is 2. The channel numbers in these base features are $16, 32, 64$, which are input channels numbers for different sub-networks. The growth rates of the 3 sub-networks are $6, 12, 24$. Moreover, for each *Fusion Block*, a compress factor of 0.25 is applied, which means that 75% of the new added channels are generated from the current sub-network and the other 25% are calculated from the previous sub-network with lower feature resolution. Furthermore, we add s transition layers for Sub-network s. E.g., we add one 3 transition layers for Sub-network 3. The Model-C-1 has six classifiers in total, and its overall architecture is illustrated in Figure 1(a).

Model-C-2: The size of four base features are $32 \times 32, 16 \times 16, 16 \times 16, 8 \times 8$. These four sub-networks corresponding to the base features have $8, 6, 4, 2$ *Conv Blocks*, respectively. Moreover, the numbers of input channels and the growth rates are $16, 32, 32, 64$ and $6, 12, 12, 24$, respectively. All *Up-Conv Layers* are substituted to *Regular Conv Layers* if the feature fusion happens between two same scales. The Model-C-2 has eight classifiers in total, and its overall architecture is illustrated in Figure 1(b).

Model-C-3: The size of four base features are $32 \times 32, 16 \times 16, 8 \times 8, 8 \times 8$. The numbers of input channels and the growth rates are $16, 16, 32, 64$ and $6, 6, 12, 24$, respectively. All *Up-Conv Layers* are substituted to *Regular Conv Layers* if the feature fusion happens between two same scales. The Model-C-3 has eight classifiers in total, and its overall architecture is illustrated in Figure 1(c).

In the experiments, the Model-C-3 (even) are evaluated under the anytime classification setting (Figure 5 of the main paper), and all three models (lg) are evaluated under the budgeted batch classification setting (Figure 6 of the main paper).

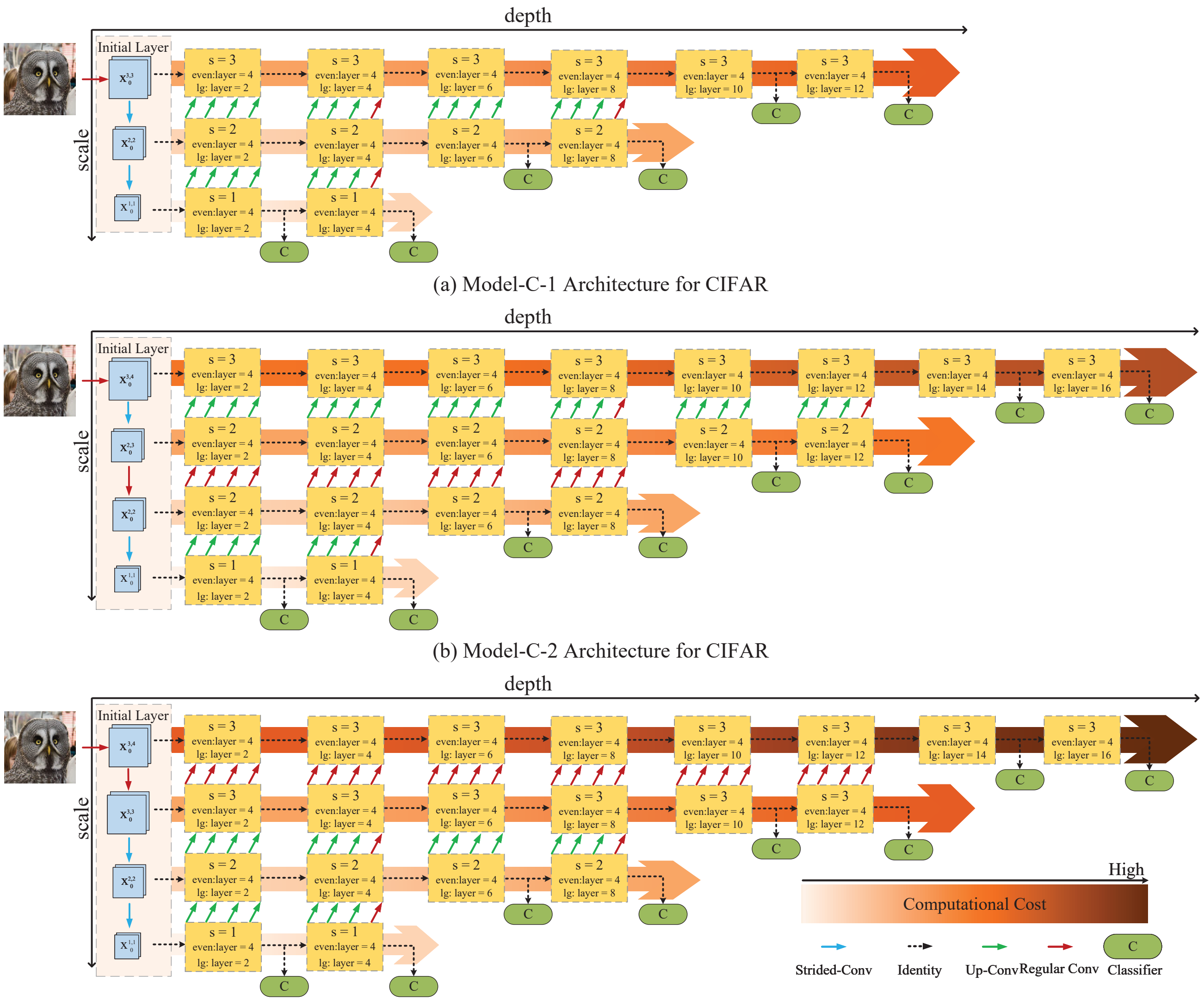
1.2. ImageNet

MSDNet: On the ImageNet, features with 4 different scales ($56 \times 56, 28 \times 28, 14 \times 14, 7 \times 7$) are used for MSDNets in our experiments. Three different MSDNets with five classifiers and different depth are evaluated. Specifically, the i^{th} classifier is attached at the $(t \times i + 3)^{th}$ layer where $i \in \{1, \cdots, 5\}$, and $t \in \{4, 6, 7\}$ is the step (number of layers) for each network block.

RANet: The same 4 feature scales are utilized for our RANets in the experiments. The spatial resolutions of the base features are $56 \times 56, 28 \times 28, 14 \times 14, 7 \times 7$, respectively. We test 2 architecture configurations as follows:

Model-I-1: Four sub-networks corresponding to the base features have $8, 6, 4, 2$ *Conv Blocks*, respectively, and the number of layer in each *Conv Block* is set to 8. Moreover, the numbers of base feature channels and the growth rates are $32, 64, 64, 128$ and $16, 32, 32, 64$. For each *Fusion Block*, compress factor of 0.25 is applied. The Model-I-1 has eight classifiers in total, and its overall architecture is illustrated in Figure 2.

Model-I-2: The architecture of the Model-I-2 is exactly the same as the Model-I-1. However, the numbers of base feature channels are $64, 128, 128, 256$.



(a) Model-C-1 Architecture for CIFAR

(b) Model-C-2 Architecture for CIFAR

(c) Model-C-3 Architecture for CIFAR

Figure 1. Architecture of RANets for CIFAR-10 and CIFAR-100.

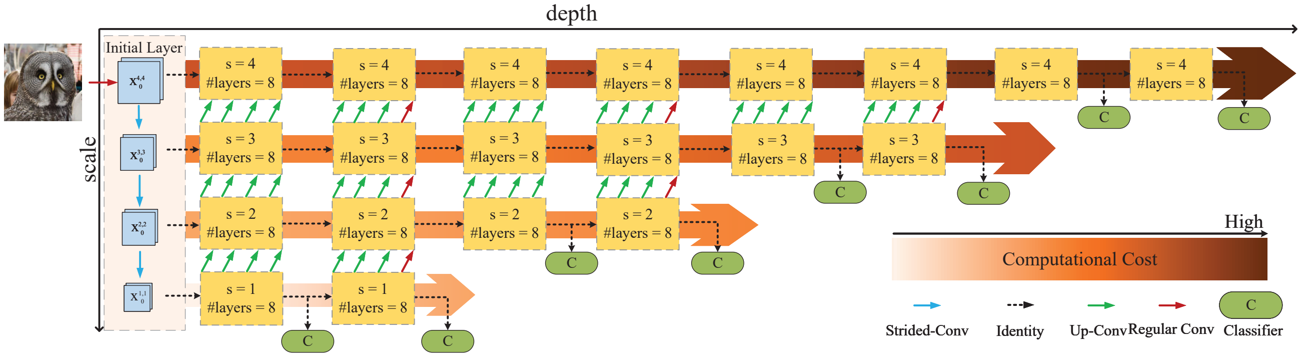


Figure 2. Architecture of RANets for ImageNet.

In the experiments, the Model-I-2 is evaluated under the anytime classification setting (Figure 5 of the main paper), and both models are evaluated under the budgeted batch classification setting (Figure 6 of the main paper).

2. Appendix B: Improved Techniques

As some training techniques for adaptive inference models with multiple exits have been proposed in [2], we further evaluated the proposed RANet and MSDNet [1] with the implementation of these improved techniques on CIFAR-100. Inline Sub-network Collaboration (ISC) and One-For-All (OFA) knowledge distillation approaches are utilized in the experiments under anytime prediction and budgeted batch classification settings. Specifically, we implement these techniques (ISC and OFA) on our **Model-C-3** and MSDNet with 8 and 10 classifiers. The results are shown in Figure 3 (anytime) and 4 (budgeted batch).

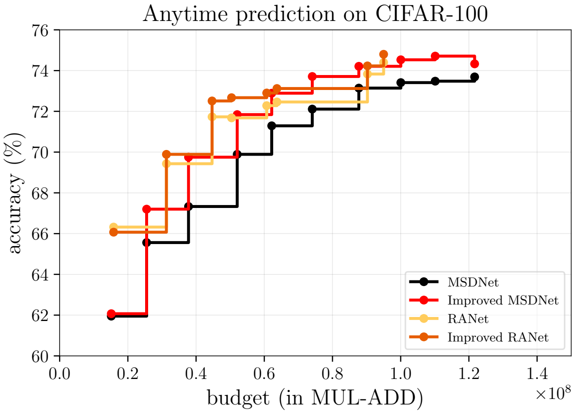


Figure 3. Accuracy (top-1) of anytime classification models as a function of average computational budget per image the on CIFAR-100, higher is better. MSDNet and RANet are trained with and without ISC and OFA techniques.

For anytime prediction, Model-C-3 (even) and MSDNet with 10 classifiers are tested. From the results, we observe that the improved RANet can outperform the improved MSDNet, especially when the budget ranges from 0.3×10^8 to 0.6×10^8 FLOPs. Moreover, the improved RANet can achieve the highest accuracy (75%) with around 0.2×10^8 less FLOPs. We further observe that the techniques (ISC and OFA) do not work well on the first classifier of the RANet.

For budgeted batch classification, the results of RANet, Model-C-3 and MSDNet with 8 classifiers are tested. From the results, we observe that the improved RANet is still superior to the improved MSDNet, especially when the budget greater than 0.3×10^8. The original RANet can outperform the improved RANet can be due to the performance dropping of the first classifiers. However, compared with

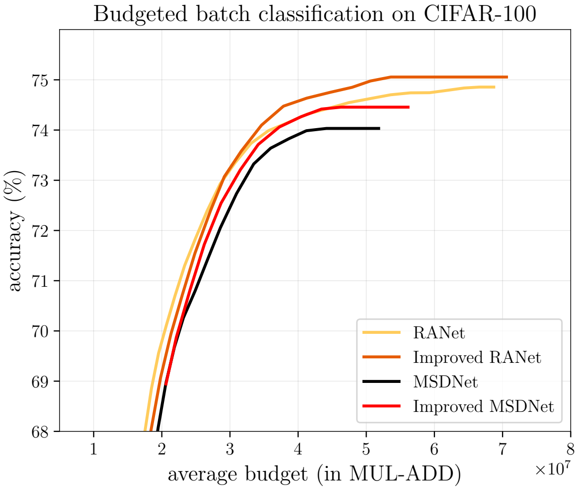


Figure 4. Accuracy (top-1) of budgeted batch classification models as a function of average computational budget per image the on CIFAR-100, higher is better. MSDNet and RANet are trained with and without ISC and OFA techniques.

MSDNet and improved MSDNet, the accuracy of improved RANet can be 1% and 0.5% higher respectively, which demonstrated the effectiveness of our RANet when implemented with the improved techniques.

References

[1] Gao Huang, Danlu Chen, Tianhong Li, Felix Wu, Laurens van der Maaten, and Kilian Q Weinberger. Multi-scale dense networks for resource efficient image classification. 2018. 3

[2] Hao Li, Hong Zhang, Xiaojuan Qi, Ruigang Yang, and Gao Huang. Improved techniques for training adaptive deep networks. In *ICCV*, 2019. 3