A1. More Implementation Details

Training and evaluation details. We use an SGD optimizer with a momentum of 0.9 and a weight decay of 10^{-4} in our experiments. And we choose the model with the best validation accuracy during the training process. Besides, we use an early weight rewinding (Frankle et al., 2019a) method (rewind to the third epoch) to help scale up the lottery ticket hypothesis in these models, except for the warmup and low variant of ResNet-20 and ResNet-56, in which the weight will be rewound to the same random initialization. For the variant of warmup, we replace the original 85 epochs (Frankle & Carbin, 2018) with 15 epochs, which does not affect the performance. The threshold of the number of forgets is set to 0 and we default to use 0.07 as the threshold for the distance between masks. Note that our baseline results are aligned with (Frankle & Carbin, 2018).

Dataset. We consider three datasets in our implementation, which can be download at `https://www.cs.toronto.edu/~kriz/cifar.html` for CIFAR-10 and CIFAR-100, and `http://cs231n.stanford.edu/tiny-imagenet-200.zip` for Tiny-ImageNet. For all three datasets, 10 percent of data from the training set are randomly split up as validation set. And we utilize random cropping and random horizontal flipping for data augmentation.

Computing infrastructures. All our experiments are conducted on Quadro RTX 6000 and Tesla V100 GPUs.

A2. More Experiment Results

A2.1. More Results of Sampling Strategy

As shown in Figure A10, our PrAC sets achieve consistent improvement compare with other sampling strategies. It indicates that our approach produces more informative pruning-aware subsets and contribute for finding high-quality winning tickets.

A2.2. More Results of Different Lottery Ticket Settings

Figure A11 reports the performance on CIFAR-100 with ResNet-56 under two additinal lottery tickets settings, **low** and **warmup**. We can observe that our methods cost 34.33% $\sim$ 38.01% training sources and achieve comparable performance, which suggests the efficiency of our PrAC lottery tickets.

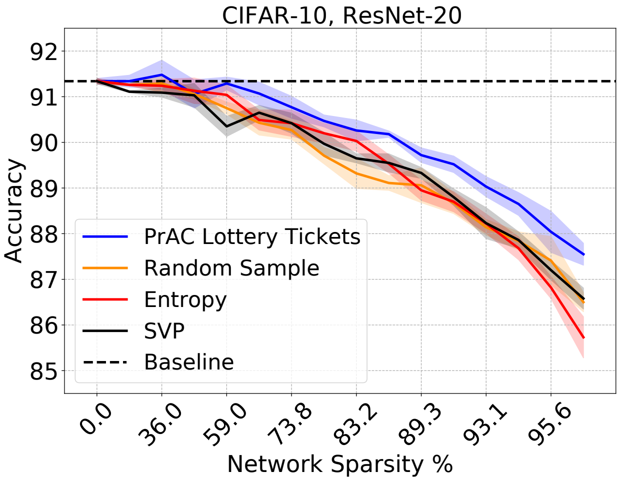

Figure A10. Comparison of our PrAC sets with other core-sets or active learning approaches.

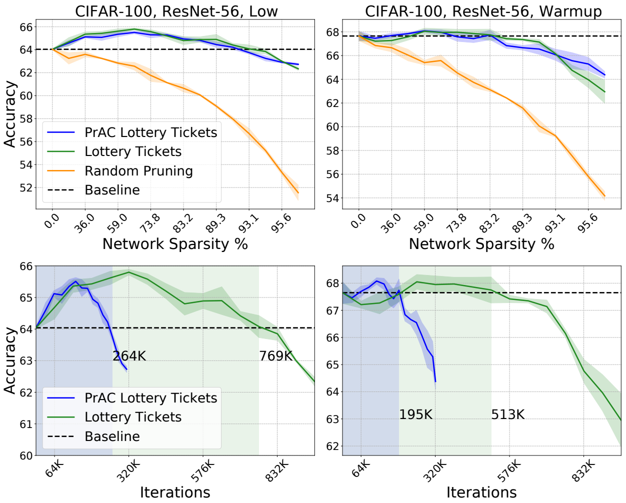

Figure A11. Testing accuracy of subnetworks at a range of sparsity levels from 0% to 96.48% (the first row) and the training iterations for finding each subnetwork (the second row) on CIFAR-100 with ResNet-56 under different lottery ticket settings. Blue, Green, Orange, and **Black** curves represent our PrAC lottery tickets, vanilla lottery tickets, random pruning and full network, respectively. The numbers within figures are the iterations used to find the subnetworks with the **same sparsity** and **comparable performance**.

A2.3. More Statistics of PrAC Sets

Table A2 contains the size of PrAC sets across different datasets and networks. On CIFAR-10 (10 classes), we locate PrAC sets with the size range from 35.32% to 37.07% of the training set, while 69.55% to 78.19% on CIFAR-100 (100 classes) and 68.23% to 75.10% on Tiny-ImageNet (200 classes). The result suggests that more data are needed to find high-quality PrAC lottery tickets for the image recognition with more classes.

A2.4. More Results of Ablation Study

The two components in the PrAC set. We conduct our data and model co-design framework with only critical examples for training (CET), named as CET lottery tickets. As

Table A2. Proportion of PrAC sets to their training set sizes of CIFAR-10, CIFAR-100 and Tiny-ImageNet

Dataset	Network	Proportion of PrAC sets
CIFAR-10	ResNet-20	36.66%
	ResNet-56	37.07%
	VGG-16	35.32%
CIFAR-100	ResNet-20	78.19%
	ResNet-56	74.94%
	VGG-16	69.55%
Tiny-ImageNet	ResNet-18	75.10%
	VGG-16	68.23%

shown in Figure A12, without the assistance of critical examples for pruning (CEP), there is a consistent performance gap between PrAC lottery tickets and CET lottery tickets. Besides, we collect the number of CET, CEP and PrAC sets in Table A3. The overlapping rate means the percentage of the overlap images between CET and CEP sets in CEP sets, (*i.e.*, $\frac{|\mathrm{CEP}|\cap|\mathrm{CET}|}{|\mathrm{CEP}|}$). We observe that as the sparsity grows, the number of CEP sets increases while the overlapping rate decreases, which indicates gradually detached distributions of critical samples during training and pruning.

Table A3. Results of the number of the identified CET, CEP and PrAC sets, as well as the overlapping rate of CEP sets during the process of our co-design framework on CIFAR-10 with ResNet-20.

Sparsity of Subnetworks	CEP	CET	PrAC	Overlapping Rate
20.00%	1501	24159	24168	99.40%
36.00%	1481	21708	21728	98.65%
48.80%	3935	19542	19838	92.48%
59.04%	3782	17674	18161	87.12%
67.23%	4723	16091	16712	86.85%
73.79%	5514	14771	16026	77.24%
79.03%	4420	14357	15202	80.88%
83.22%	4602	13909	14880	78.90%
86.58%	5391	13741	14980	77.02%
89.26%	5360	14168	15376	77.46%
91.41%	5098	14365	15247	82.70%
93.13%	5840	14553	15804	78.58%
94.50%	5728	14959	16062	80.74%
95.60%	6370	15290	16360	83.20%
96.48%	6616	15369	16499	82.92%

Relative similarity between PrAC LT and LT. We evaluate the overlap degree in sparsity patterns with relative similarity (*i.e.*, $\frac{m_i \cap m_j}{m_i \cup m_j}$), where m_i and m_j are the sparsity masks of identified subnetworks. We keep the same random initialization for PrAC lottery tickets and two independent runs of vanilla lottery tickets. Figure A13 shows that as the sparsity grows, subnetworks share fewer sparsity patterns. And the relative similarity between PrAC lottery tickets and lottery tickets are slightly smaller than between two different runs of lottery tickets, which indicates the non-trivial difference between sparse masks of PrAC LT and LT.

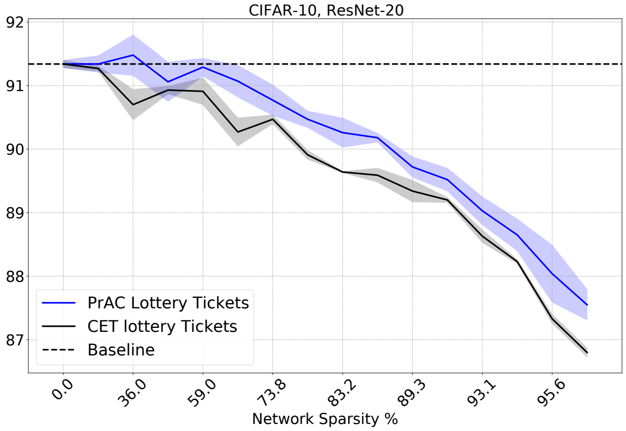

Figure A12. Comparison of PrAC lottery tickets with CET lottery tickets on CIFAR-10 with ResNet-20. Each curve contains the mean and standard deviation of testing accuracy of subnetworks at different sparsity levels.

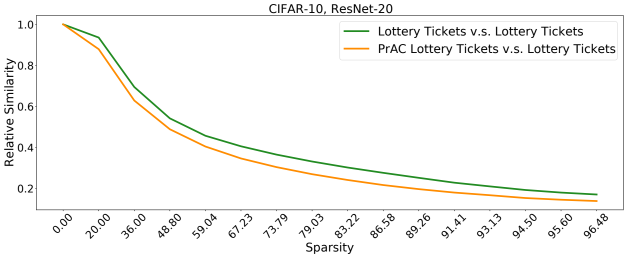

Figure A13. Results of the relative mask similarity on CIFAR-10 with ResNet-20. Green and Orange represents the relative similarity between two independent runs of vanilla lottery tickets, and the one between PrAC lottery tickets and vanilla lottery tickets. We adopt the same random initialization for identifying these three groups of subnetworks.

Lottery tickets with subsets of random sampling To investigate that how many examples of random sampling can match the performance of our PrAC subsets in terms of locating subnetworks, we conduct an ablation study on CIFAR-10 with ResNet-20 and record the results in Figure A14. We can observe that nearly 70% data are needed for random subsets to match the performance of our PrAC sets, which only contain 37% $\sim$ 54% data.

Table A4. Results of test accuracy of identified subnetworks with respect to the threshold for the number of forgets on CIFAR-10 with ResNet-20. We select subnetworks with the same sparsity of 16.78%, which is the maximum sparsity of subnetworks identified by PrAC subsets ($\mathcal{E}_{\mathrm{F}} = 0$) have **comparable performance**.

$\mathcal{E}_{\mathrm{F}}$	0	2	4	6	8	10
Accuracy (%)	91.05	90.43	90.35	89.32	89.08	88.79
PrAC	19748	14992	12536	11141	11152	10338

The threshold for the number of forgets Table A4 records the test accuracy and the size of PrAC subsets un-

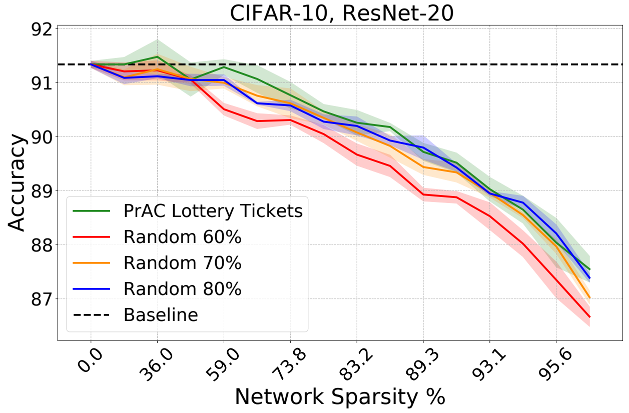

Figure A14. Comparison of the quality of subnetworks identified by our PrAC subsets and subsets from random sampling on CIFAR-10 with ResNet-20. We keep the size of random subsets consistent during the whole IMP process, ranging from $60\% \sim 80\%$.

der different threshold for the number of forgets. We can observe that both the test accuracy and the size of PrAC decrease as the threshold rises. Thus we choose $\mathcal{E}_{\mathrm{F}} = 0$ in our implementation.

A2.5. More Visualization and Analyses

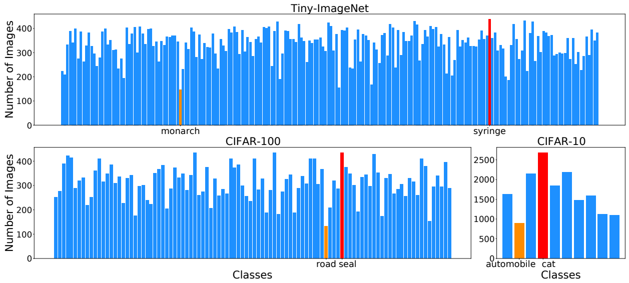

Figure A15. The class-wise ratios of images in PrAC sets on CIFAR-10/100 and Tiny-ImageNet, respectively. Red and Orange represent the classes with maximum and minimum images.

Figure A15 demonstrates the class-wise ratios of images in PrAC set, from which we can find that the number of images from different classes are in the same order. This balanced distribution of PrAC set's classes may provide possible insights on the effectiveness of PrAC sets, with respect to locating critical subnetworks, i.e., PrAC tickets, with satisfying performance.

A2.6. Additional Results of Forgetting Statistics in LT

Figure A16 shows the distribution of training data's forgetting times at different sparsity from 0% to 96.48% on CIFAR-10 with ResNet-20. We consider three pruning methods: Basic iterative magnitude pruning (IMP) (Han et al., 2015), vanilla lottery tickets (LT) (Frankle & Carbin, 2018), and random tickets (RT). IMP fine-tune the subnetworks directly after pruning while LT rewinds the weight to the same initialization and RT reinitializes the subnetworks before fine-tuning. We can observe that as the sparsity increases, for IMP and LT, the number of unforgettable images first increases and then decreases, while the one for RT consistently decreases. Besides, the maximum number of forgetting times grows as the sparsity becomes larger.

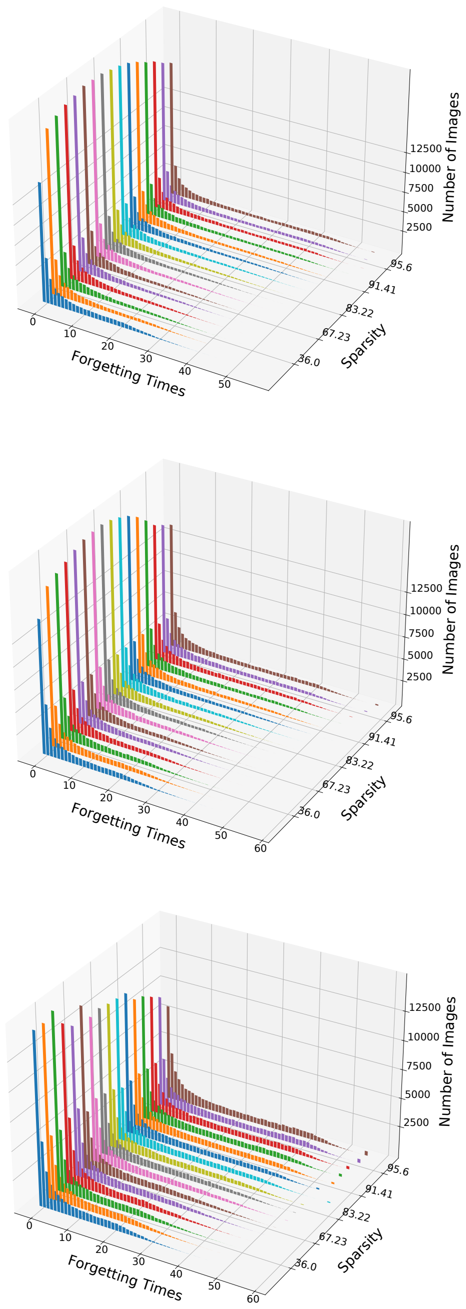

Figure A16. Visualization of the forgetting statistics of subnetworks at different sparsity from 0% to 96.48% on CIFAR-10 with ResNet-20 when training with full data. *Top:* Basic iterative magnitude pruning (fine-tune after pruning). *Middle:* vanilla lottery tickets. *Bottom:* random tickets.