A. Supplementary material

We provide additional results / observations regarding temperature annealing (Section A.1), the choice of teachers (Section A.2), the semi-supervised setting with very few labels (Section A.3), inference times for different modes and exits (Section A.4), and the anytime-mode (Section A.5).

A.1. Temperature annealing

Expanding on the observations of Section 4.4, we study the effect of temperature annealing on final accuracy. Figure 8 (same as Figure 7 in the main text) shows the accuracy curves of five MSDNets trained by distillation, each with a different temperature setting. We compare the proposed annealing scheme (green) to training with constant temperature, $T \in \{1.0, 2.0, 3.0, 4.0\}$. The figure shows that while using no temperature at all ($T = 1.0$) leads to significant accuracy drops, as long as the temperature is “high enough”, its exact value seems to matter little for the final model’s accuracy. Still, the proposed annealing scheme performs as well as, or better than any choice of a constant temperature, and has the advantage of being easier to tune.

A.2. Choice of teachers

We performed exploratory experiments with a different choice of teacher sets $\mathcal{T}$. In particular, we let each exit learn from all later exits, $\mathcal{T}(m) = \{m+1, \ldots, M\}$ for $m < M$ and $\mathcal{T}(M) = \emptyset$. The intuition behind this choice is that learning from an ensemble of good exits might be better than learning from a single good exit. However, as the results show, this turns out not to be the case.

Figure 9 shows the accuracy curve of a model trained by distillation from all later exits (yellow), as well as curves for distillation from only the last exit (green) and exit-wise training (blue) for comparison. We observe that the two teacher-set choices yield models of very similar accuracy, so there seems to be little benefit in adopting more complicated teacher-student setups.

A.3. Semi-supervised distillation with few labels

In the main text, we show that using additional unlabelled data with distillation-based training may improve accuracy. However, the gains are relatively modest, and only significant for the late exits in the case of ImageNet(100) and the early exits in the case of CIFAR(150) and CIFAR(250). One could speculate why the gains are not larger, and a potential explanation would be that in the considered settings, the amount of labeled data is too high. To test this hypothesis, we ran the experiment from Section 4.1 on CIFAR(80), i.e. using only 30 labelled images per class for training and 50 for validation. However, the gains remain rather small (see Figure 10), at most 1%, similar to the case of CIFAR(150) and CIFAR(250).

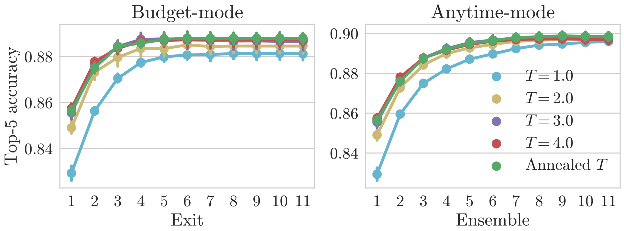

Figure 8: Top-5 accuracy of five models trained by distillation, each with a different temperature setting, on CIFAR(250). Results shown for different computational budgets in both the budget-mode (left) and the anytime-mode (right).

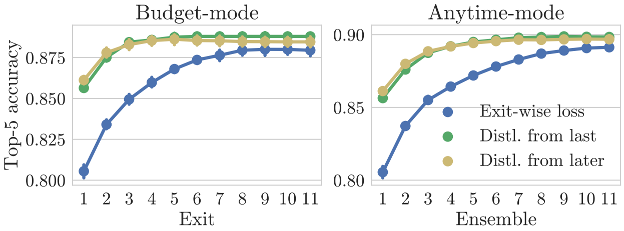

Figure 9: Top-5 accuracy of a model trained by distillation from all later exits (yellow) vs. trained by distillation from the last exit (green) vs. trained by the exit-wise loss (blue), on CIFAR(250). Results shown for different computational budgets in both the budget-mode (left) and the anytime-mode (right).

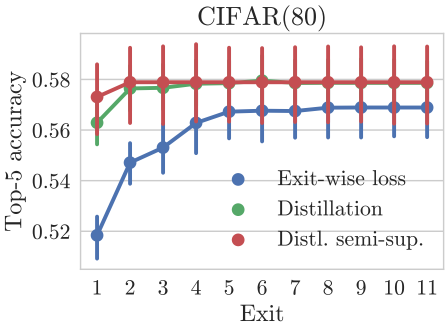

Figure 10: Top-5 accuracy as a function of computational budget (denominated in available exits). MSDNet trained by the exit-wise loss (blue) vs. trained by distillation (green) vs. trained by semi-supervised distillation (red) on CIFAR(80).

A.4. Example inference times

To give the reader a sense of the efficiency gains achievable by anytime inference, we provide some example inference times for CIFAR MSDNet (Table 3) and ImageNet MSDNet (Table 4), for the two inference modes (budget-mode and anytime-mode) and different exits. We made no effort to optimize these timings. They are obtained by simply running our code, without any changes, and measuring the time elapsed. We expect the results to generalize qualitatively to different hardware and implementations, though the exact numbers are likely to vary.

A.5. Complete results for anytime-mode

For the reader's convenience, we collect here all results pertaining to the anytime-mode. We report the accuracy of an MSDNet trained by the exit-wise loss, by distillation, and by distillation using additional unlabeled data. Results for ImageNet are shown in Figure 4 (reprinted from the main text), and in Table 5. Results for CIFAR100 are shown in Figure 5 (reprinted from the main text), and in Table 6. The results are similar to those for the budget-mode: distillation-based training clearly outperforms exit-wise training. See Section 4.3 in the main text for a more detailed discussion.

	CPU timings [s]		GPU timings [s]	
	Budget-mode	Anytime-mode	Budget-mode	Anytime-mode
Exit / Ensemble 1	0.024 ± 0.022	0.026 ± 0.029	0.007 ± 0.000	0.007 ± 0.000
Exit / Ensemble 2	0.034 ± 0.029	0.043 ± 0.037	0.011 ± 0.000	0.011 ± 0.000
Exit / Ensemble 3	0.043 ± 0.030	0.062 ± 0.040	0.015 ± 0.000	0.015 ± 0.000
Exit / Ensemble 4	0.051 ± 0.035	0.081 ± 0.049	0.018 ± 0.000	0.019 ± 0.000
Exit / Ensemble 5	0.062 ± 0.041	0.099 ± 0.055	0.022 ± 0.001	0.023 ± 0.001
Exit / Ensemble 6	0.067 ± 0.034	0.121 ± 0.064	0.025 ± 0.001	0.027 ± 0.001
Exit / Ensemble 7	0.071 ± 0.046	0.139 ± 0.066	0.029 ± 0.001	0.032 ± 0.001
Exit / Ensemble 8	0.085 ± 0.035	0.164 ± 0.078	0.032 ± 0.001	0.035 ± 0.001
Exit / Ensemble 9	0.093 ± 0.035	0.169 ± 0.085	0.036 ± 0.001	0.039 ± 0.001
Exit / Ensemble 10	0.103 ± 0.040	0.196 ± 0.078	0.039 ± 0.001	0.043 ± 0.001
Exit / Ensemble 11	0.105 ± 0.038	0.219 ± 0.085	0.041 ± 0.001	0.045 ± 0.001

Table 3: Inference times for the CIFAR MSDNet operating either in the budget-mode or the anytime-mode, evaluated either on CPU or GPU. We report the mean ± stdev over 1000 runs.

	CPU timings [s]		GPU timings [s]	
	Budget-mode	Anytime-mode	Budget-mode	Anytime-mode
Exit / Ensemble 1	0.159 ± 0.048	0.229 ± 0.079	0.027 ± 0.001	0.027 ± 0.002
Exit / Ensemble 2	0.339 ± 0.099	0.373 ± 0.094	0.038 ± 0.001	0.038 ± 0.001
Exit / Ensemble 3	0.471 ± 0.131	0.519 ± 0.118	0.048 ± 0.001	0.048 ± 0.002
Exit / Ensemble 4	0.559 ± 0.138	0.647 ± 0.117	0.055 ± 0.001	0.056 ± 0.001
Exit / Ensemble 5	0.665 ± 0.125	0.703 ± 0.145	0.057 ± 0.001	0.059 ± 0.001

Table 4: Inference times for the ImageNet MSDNet operating either in the budget-mode or the anytime-mode, evaluated either on CPU or GPU. We report the mean ± stdev over 1000 runs.

	ImageNet(100)			ImageNet(300)		
	Exit-wise loss	Distillation	Distl. semi-sup.	Exit-wise loss	Distillation	Distl. semi-sup.
Ensemble 1	64.4 ± 0.4	**68.1 ± 0.5**	**68.1 ± 0.4**	79.5 ± 0.2	**82.3 ± 0.2**	**82.3 ± 0.3**
Ensemble 2	67.7 ± 0.3	**69.7 ± 0.6**	**69.9 ± 0.2**	82.3 ± 0.1	**84.0 ± 0.2**	**84.1 ± 0.5**
Ensemble 3	69.2 ± 0.3	**70.3 ± 0.6**	**70.6 ± 0.1**	83.6 ± 0.2	**84.7 ± 0.3**	**84.8 ± 0.4**
Ensemble 4	69.9 ± 0.3	70.5 ± 0.6	**70.8 ± 0.2**	84.3 ± 0.2	**84.9 ± 0.3**	**85.0 ± 0.5**
Ensemble 5	70.2 ± 0.4	70.5 ± 0.7	70.8 ± 0.2	84.6 ± 0.2	85.0 ± 0.3	85.1 ± 0.5

	ImageNet(500)			ImageNet(full)	
	Exit-wise loss	Distillation	Distl. semi-sup.	Exit-wise loss	Distillation
Ensemble 1	83.4 ± 0.2	**85.6 ± 0.1**	**85.4 ± 0.1**	87.8 ± 0.2	**88.8 ± 0.1**
Ensemble 2	86.3 ± 0.1	**87.3 ± 0.2**	**87.1 ± 0.3**	90.2 ± 0.1	**90.5 ± 0.1**
Ensemble 3	87.6 ± 0.2	**88.0 ± 0.2**	87.9 ± 0.3	91.3 ± 0.1	91.3 ± 0.1
Ensemble 4	88.1 ± 0.2	88.3 ± 0.3	88.2 ± 0.4	**91.9 ± 0.1**	91.6 ± 0.1
Ensemble 5	88.5 ± 0.2	88.4 ± 0.3	88.4 ± 0.4	**92.2 ± 0.1**	91.8 ± 0.1

Table 5: Top-5 accuracy in % (mean ± 1.96 stderr) of first-m-exits ensembles ($m = 1, \dots, 5$) trained by the exit-wise loss vs. trained by distillation vs. trained by semi-supervised distillation on ImageNet ILSVRC2012 with 100, 300, 500 or all available (≥ 700) training images per class.

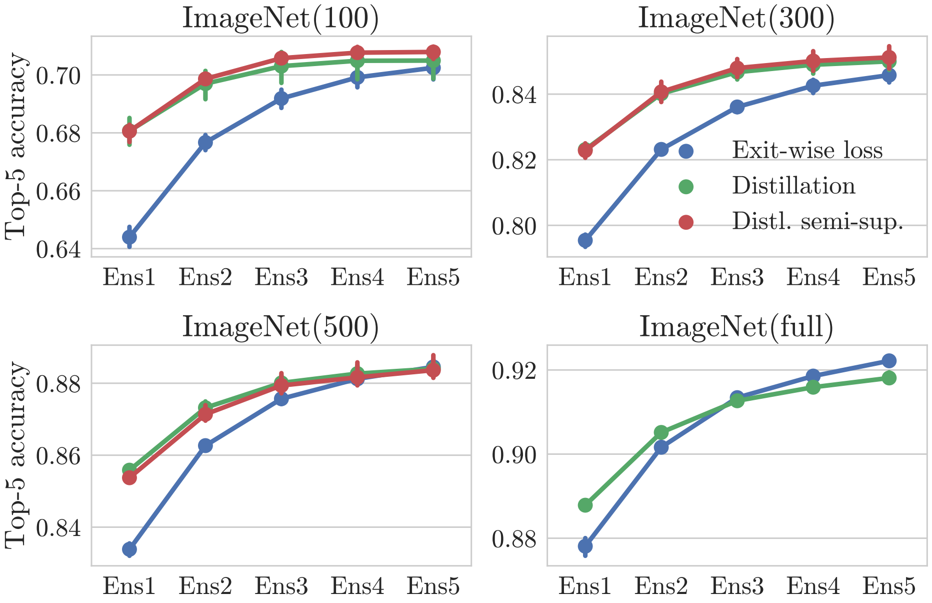

Figure 4: Top-5 accuracy of first-m-exits ensembles ($m = 1, \dots, 5$) trained by the exit-wise loss (blue) vs. trained by distillation (green) vs. trained by semi-supervised distillation (red) on ImageNet ILSVRC2012 with 100, 300, 500 or all available (≥ 700) training images per class. (Figure repeated from page 8.)

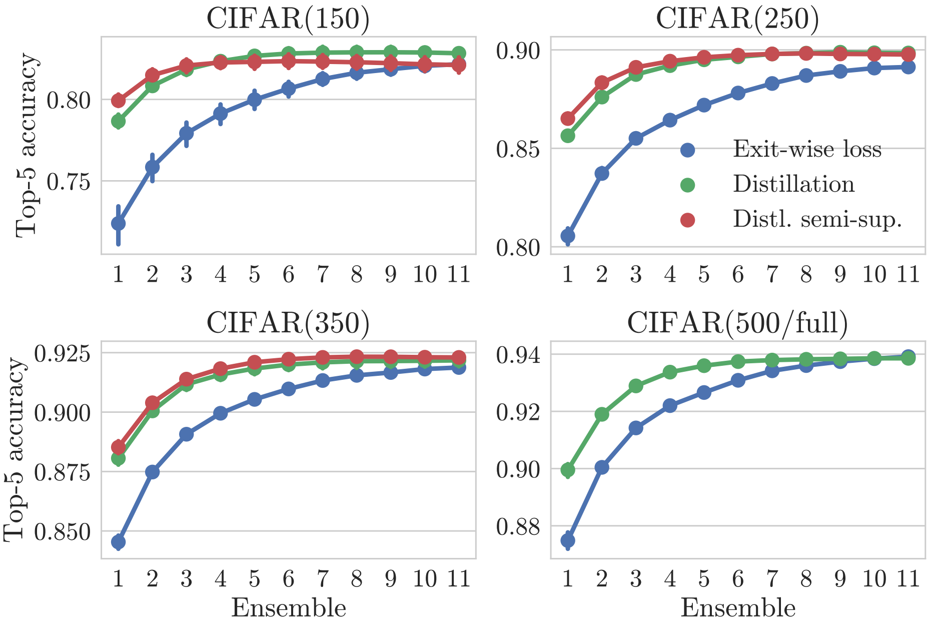

Figure 5: Top-5 accuracy of first-m-exits ensembles ($m = 1, \dots, 11$) trained by the exit-wise loss (blue) vs. trained by distillation (green) vs. trained by semi-supervised distillation (red) on CIFAR100 with 150, 250, 350 or 500 images per class. (Figure repeated from page 8.)

	CIFAR(150)			CIFAR(250)		
	Exit-wise loss	Distillation	Distl. semi-sup.	Exit-wise loss	Distillation	Distl. semi-sup.
Ensemble 1	72.4 ± 1.3	**78.7 ± 0.4**	**79.9 ± 0.4**	80.6 ± 0.4	**85.6 ± 0.2**	**86.5 ± 0.3**
Ensemble 2	75.8 ± 0.9	**80.8 ± 0.3**	**81.5 ± 0.4**	83.7 ± 0.3	**87.6 ± 0.2**	**88.3 ± 0.3**
Ensemble 3	77.9 ± 0.8	**81.8 ± 0.3**	**82.1 ± 0.4**	85.5 ± 0.2	**88.7 ± 0.2**	**89.1 ± 0.2**
Ensemble 4	79.1 ± 0.7	**82.3 ± 0.3**	**82.3 ± 0.4**	86.4 ± 0.2	**89.2 ± 0.2**	**89.4 ± 0.2**
Ensemble 5	80.0 ± 0.6	**82.7 ± 0.3**	**82.3 ± 0.4**	87.2 ± 0.2	**89.5 ± 0.2**	**89.6 ± 0.2**
Ensemble 6	80.7 ± 0.5	**82.8 ± 0.4**	**82.3 ± 0.4**	87.8 ± 0.2	**89.6 ± 0.2**	**89.7 ± 0.2**
Ensemble 7	81.2 ± 0.4	**82.9 ± 0.4**	**82.3 ± 0.4**	88.3 ± 0.2	**89.8 ± 0.2**	**89.8 ± 0.2**
Ensemble 8	81.6 ± 0.4	**82.9 ± 0.4**	82.3 ± 0.4	88.7 ± 0.2	**89.8 ± 0.2**	**89.8 ± 0.2**
Ensemble 9	81.9 ± 0.3	**82.9 ± 0.3**	82.2 ± 0.4	88.9 ± 0.2	**89.9 ± 0.2**	**89.8 ± 0.2**
Ensemble 10	82.1 ± 0.3	**82.9 ± 0.4**	82.2 ± 0.5	89.1 ± 0.2	**89.9 ± 0.2**	**89.8 ± 0.2**
Ensemble 11	82.2 ± 0.3	82.8 ± 0.3	82.1 ± 0.5	89.1 ± 0.2	**89.8 ± 0.2**	**89.8 ± 0.2**

	CIFAR(350)			CIFAR(500)	
	Exit-wise loss	Distillation	Distl. semi-sup.	Exit-wise loss	Distillation
Ensemble 1	84.5 ± 0.3	**88.1 ± 0.3**	**88.5 ± 0.3**	87.5 ± 0.3	**90.0 ± 0.2**
Ensemble 2	87.5 ± 0.2	**90.0 ± 0.2**	**90.4 ± 0.2**	90.0 ± 0.1	**91.9 ± 0.2**
Ensemble 3	89.1 ± 0.2	**91.2 ± 0.2**	**91.4 ± 0.2**	91.4 ± 0.1	**92.9 ± 0.1**
Ensemble 4	90.0 ± 0.2	**91.6 ± 0.2**	**91.8 ± 0.2**	92.2 ± 0.1	**93.4 ± 0.1**
Ensemble 5	90.5 ± 0.2	**91.8 ± 0.3**	**92.1 ± 0.2**	92.7 ± 0.1	**93.6 ± 0.1**
Ensemble 6	91.0 ± 0.1	**92.0 ± 0.3**	**92.2 ± 0.1**	93.1 ± 0.1	**93.7 ± 0.1**
Ensemble 7	91.3 ± 0.1	**92.1 ± 0.2**	**92.3 ± 0.2**	93.4 ± 0.1	**93.8 ± 0.1**
Ensemble 8	91.5 ± 0.1	**92.1 ± 0.3**	**92.3 ± 0.2**	93.6 ± 0.1	**93.8 ± 0.1**
Ensemble 9	91.7 ± 0.1	**92.1 ± 0.3**	**92.3 ± 0.2**	93.7 ± 0.1	93.8 ± 0.1
Ensemble 10	91.8 ± 0.1	92.2 ± 0.3	**92.3 ± 0.2**	93.8 ± 0.1	93.9 ± 0.1
Ensemble 11	91.9 ± 0.1	92.2 ± 0.2	**92.3 ± 0.2**	93.9 ± 0.1	93.9 ± 0.1

Table 6: Top-5 accuracy in % (mean ± 1.96 stderr) of first-m-exits ensembles trained by the exit-wise loss vs. trained by distillation vs. trained by semi-supervised distillation on CIFAR100 with 150, 250, 350 or 500 images per class.