A. Training Details

Datasets. The statistics of the recognition benchmarks used in our paper are shown in Table 10.

Table 10. Statistics of the recognition benchmarks used in the paper.

Datasets	# Category	# Training	# Testing
CIFAR-10	10	50,000	10,000
CIFAR-100	100	50,000	10,000
Flowers	102	2,040	6,149
Pets	37	3,680	3,669
DTD	47	3,760	1,880

Training details for SSL methods. Training details for SimCLR, BYOL, SimSiam and our US3L on CIFAR-100 are shown in Table 11.

Table 11. Training details for SimCLR, BYOL, SimSiam and our US3L on CIFAR-100 in Table 2, Table 7 and Table 8. τ denotes the temperature parameter, and m denotes the momentum coefficient for the momentum network.

Method	Settings								
	bs	lr	wd	epochs	optimizer	lr sche.	τ	m	dim
SimSiam	512	0.1	5e-4	400	SGD	cosine	-	-	2048
SimCLR	512	0.5	1e-4	400	SGD	cosine	0.5	-	2048
BYOL	512	0.1	5e-4	400	SGD	cosine	-	0.99	2048
Ours	512	0.5	1e-4	400	SGD	cosine	0.5	0.99	2048

Training details for linear evaluation and fine-tuning. For ImageNet linear evaluation, we follow the same settings in [9]. For linear evaluation on other datasets, we train for 100 epochs with lr initialized to 30.0, which is divided by 10 at the 60-th and 80-th epoch.

Source codes. We promise that all codes will be made publicly available upon acceptance of the paper.

B. More Results

B.1. Transfer Results for ResNet-18

We plot the downstream object detection performance on Pascal VOC07&12 for ResNet-18 FPN in Fig. 4. Moreover, we present the transfer results on downstream recognition benchmarks for ResNet-18 in Table 12. The results show that our method is also effective for ResNet-18 when transferring to downstream object detection and recognition tasks.

B.2. Ablation Studies of Dynamic Sampling

In this subsection, we conduct ablation studies of our dynamic sampling strategy and show how we successfully reduced the sampling number s from 4 to 3 by using dynamic sampling while improving accuracy. We also compute the expected total forward number for T iterations. Take our dynamic sampling as an example, we train the largest model only in the first $\frac{T}{4}$ iterations and sample three sub-networks

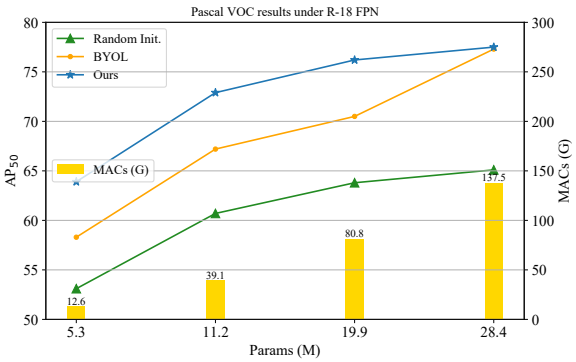


Figure 4. Transfer results on Pascal VOC 07&12 under R18-FPN.

Table 12. Transfer results on recognition benchmarks under linear evaluation. 'C-10/100' denotes 'CIFAR-10/100'.

Net	Width	Params	MACS	Method	Linear Accuracy (%)				
					C-10	C-100	Flowers	Pets	Dtd
R-18	1.0x	11.69M	1.82G	BYOL	76.6	48.6	83.0	71.1	64.8
				Ours	**77.9**	**52.6**	**84.9**	**71.2**	**65.2**
	0.75x	6.68M	1.05G	BYOL	76.4	48.1	82.6	**71.0**	**63.7**
				Ours	**76.7**	**49.0**	**83.5**	**71.0**	**63.7**
	0.5x	3.06M	0.49G	BYOL	74.6	46.9	81.4	**67.6**	61.7
				Ours	**75.2**	**47.4**	**82.0**	67.3	**62.6**
	0.25x	0.83M	0.14G	BYOL	67.0	41.2	75.5	57.6	56.4
				Ours	**67.8**	**41.4**	**77.0**	**60.6**	**56.6**

in the last $\frac{3T}{4}$ iterations, hence the expected total forward number is:

$$1 \times \frac{T}{4} + 3 \times \frac{3T}{4} = 2.5T\,. \tag{12}$$

As shown in Table 13, our dynamic sampling strategy achieves the best accuracy-efficiency trade-off. Notice that we also investigate the two components in our dynamic sampling strategy separately and we can clearly see that 'max first' reduces the training overhead (case 4) and 'gradually reduce' improves the accuracy (case 3).

B.3. Hyper-parameter Studies of G and α

In this subsection, we study the choice of hyper-parameters G and α in our group regularization (Eq. (5)) in Table 14. Notice that when $\alpha = 0$, group regularization is equivalent to the standard L_2 normalization (case 1). We used $G = 8$ and $\alpha = 0.05$ in the paper. We also present the max decay fraction $G \times \alpha$ (i.e., the decay rate for the last group). Table 14 shows that we can achieve the best results when $G \times \alpha = 40\%$ (case 1 $\sim$ case 5). Then we keep $G \times \alpha = 40\%$ and change G to 4 or 16. In terms of hyper-parameter G, we can see that $G = 8$ (case 3) outperforms $G = 4$ (case 6) and accuracy is saturated and will not continue to increase beyond 8 ($G = 16$, case 7). It is worth noting that if we set α to a negative value which goes against our motivation (case 9), we will no longer see performance gains for large sub-networks as before. The results here further validate the effectiveness of our method and the

Table 13. Ablation studies of the sampling strategy under ResNet-18 on CIFAR-100. T denotes the total number of iterations. 'Max first' denotes whether to train the largest network only in early epochs. 'Gradually reduce' denotes whether to gradually reduce the width of the smallest network.

Case	Sampling number s	Expected forward number	Dynamic Sampling		Linear Accuracy (%)								
			Max first	Gradually reduce	1.0x	0.9x	0.8x	0.7x	0.6x	0.5x	0.4x	0.3x	0.25x
1	4	$4T$	×	×	68.1	67.4	67.0	66.3	65.3	64.4	62.7	60.8	59.9
2	3	$3T$	×	×	67.7	67.2	66.5	66.0	65.1	64.3	62.5	60.5	59.6
3	3	$3T$	×	✓	68.5	67.9	**67.2**	66.4	65.3	64.5	62.6	**61.0**	**60.2**
4	3	$\mathbf{2.5}T$	✓	×	67.8	67.6	66.8	66.2	65.3	64.5	63.0	60.6	59.8
5	3	$\mathbf{2.5}T$	✓	✓	**68.6**	**68.1**	**67.2**	**66.6**	**65.5**	**64.6**	**62.8**	60.7	59.9

correctness of our analysis in the paper.

B.4. Group Regularization is Tailored for US-Net

In this subsection, we will demonstrate that our group regularization strategy is tailored for US-Net and our analysis in the paper is valid. We apply group regularization to common SSL methods which train each model individually. As shown in Table 15, the introduction of group regularization does not bring improvements for BYOL and SimCLR, which are individually trained. It shows that the group regularization is tailored for US-Net, and the improvement is due to our unique design rather than factors such as hyper-parameters.

B.5. Our US3L Can Run at Arbitrary Width

Note that we only reported the results of width at [1.0, 0.9, 0.8, 0.7, 0.6, 0.5, 0.4, 0.3, 0.25]x for CIFAR-100 and [1.0, 0.75, 0.5, 0.25]x for ImageNet due to limited space. Actually, the pretrained model of our US3L can run at any width within the predefined width range, by only training once. As a supplement, we present the results of more widths on CIFAR-100 in Table 16 and we can see that our pretrained model can achieve a good accuracy-efficiency trade-off.

B.6. Figures

As a supplement to Table 2 in the paper, we plot the results here to more intuitively see the advantages of our method. We compare with individually trained methods in Fig. 5 and compare with the US-Net baseline in Fig. 6.

B.7. Ablation Studies of Loss Design

We present more ablation results of loss design here in Table 17, as a supplement to Table 7 in the paper. We look more closely at the 'Asymmetric Distill Head' column, which indicates whether to use an additional head for distillation. Notice that there is already an asymmetrical head itself in MSE-based methods like SimSiam and BYOL. So 'Share' refers to sharing the asymmetrical head, and 'New' refers to distillation using a brand new head.

C. More Analysis

Lemma C.1 *$s = 3$ is the theoretical minimum number of samples for US-Net [31].*

Proof. First, from [31] we know sandwich rules: Performances at all widths are bounded by the performance of the model at the smallest and the largest width. In other words, optimizing the lower and upper bounds of performance can implicitly optimize all sub-networks in a US-Net. To optimize for arbitrary widths, we need at least one randomly sampled width per iteration, except for the largest and smallest sub-networks. In conclusion, $s = 3$ is the theoretical minimum number of samples for US-Net. □

Table 14. Hyper-parameter studies of G and α in group regularization under ResNet-18 on CIFAR-100.

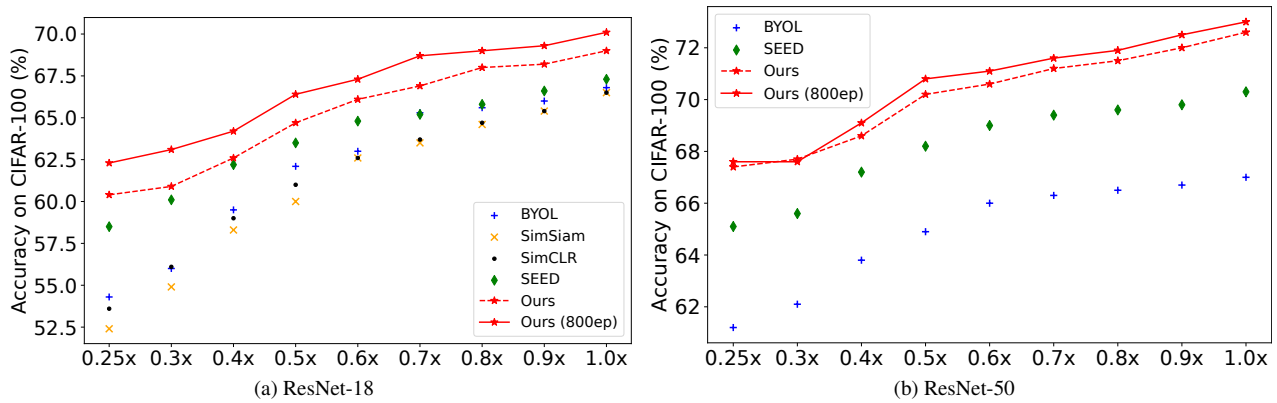
Case	Max decay fraction $G \times \alpha$	Group number G	Decay rate α	Linear Accuracy (%)										
				1.0x	0.9x	0.8x	0.7x	0.6x	0.5x	0.4x	0.3x	0.25x	Avg.	1.0x diff
1	0%	-	0	67.7	67.2	66.5	66.0	65.1	64.3	62.5	60.5	59.6	64.4	+0.0%
2	20%	8	0.025	68.0	67.4	66.3	66.0	65.2	64.0	62.6	61.0	60.3	64.5	+0.3%
3	40%	8	0.05	68.6	67.8	67.3	66.4	65.5	64.4	63.1	60.9	60.1	**64.9**	**+0.9%**
4	60%	8	0.075	68.2	67.6	66.9	66.1	65.3	64.2	62.7	61.0	60.2	64.7	+0.5%
5	80%	8	0.1	67.7	67.0	66.7	66.4	65.8	64.2	62.8	61.3	60.5	64.7	+0.0%
6	40%	4	0.1	68.0	67.4	66.6	66.1	65.1	63.8	63.0	61.6	60.6	64.7	+0.3%
7	40%	16	0.025	68.7	67.6	67.3	66.5	65.5	64.7	62.8	61.3	60.1	**64.9**	**+1.0%**
8	80%	16	0.05	67.9	67.3	66.7	66.3	65.5	64.2	63.1	61.1	60.5	64.7	+0.2%
9	-40%	8	-0.05	67.4	67.0	66.3	65.9	64.7	64.0	62.8	61.2	60.1	64.4	-0.3%

Table 15. Effect of group regularization in BYOL and SimCLR under ResNet-18 on CIFAR-100. Our group regularization strategy is tailored for US-Net.

Method	Model Type	Once Training	Group Regularization	Linear Accuracy (%)								
				1.0x	0.9x	0.8x	0.7x	0.6x	0.5x	0.4x	0.3x	0.25x
BYOL	individual	×	×	66.8	66.0	65.6	65.3	63.0	62.1	59.5	56.0	54.3
			✓	66.0	66.2	65.6	64.4	63.4	61.8	59.3	56.1	54.0
SimCLR	individual	×	×	66.5	65.4	64.7	63.7	62.6	61.0	59.0	56.1	53.6
			✓	65.7	65.0	65.0	63.3	62.6	61.0	58.8	56.0	53.4
Ours	US-Net	✓	×	67.7	67.2	66.5	66.0	65.1	64.3	62.5	60.5	59.6
			✓	**68.6**	**67.8**	**67.3**	**66.4**	**65.5**	**64.4**	**63.1**	**60.9**	**60.1**

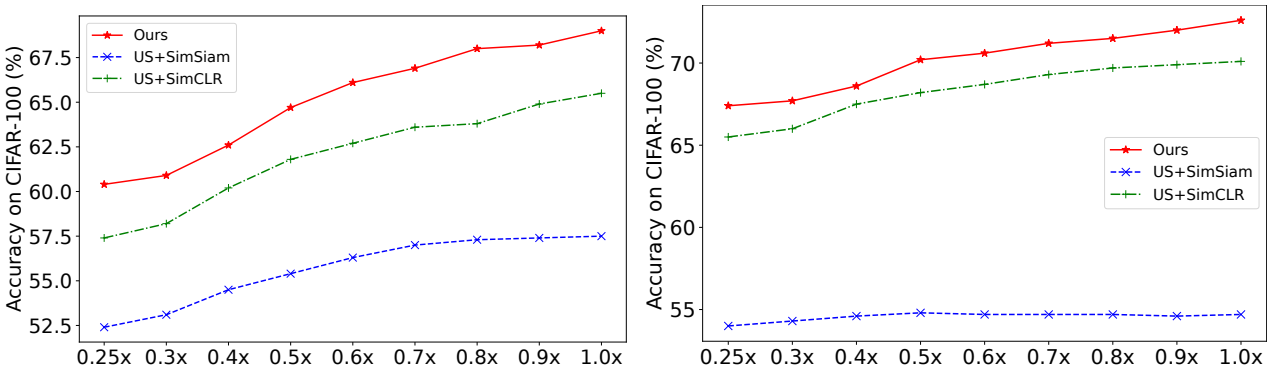
Table 16. Results of our US3L method at different widths under ResNet-18 and ResNet-50 on CIFAR-100. Our US3L can run at arbitrary width and we only reported partial results as a representative in the paper due to limited space.

Method	Backbone	Linear Accuracy (%)																
		1.0x	0.95x	0.9x	0.85x	0.8x	0.75x	0.7x	0.65x	0.6x	0.55x	0.5x	0.45x	0.4x	0.35x	0.3x	0.275x	0.25x
Ours (800ep)	R-18	70.1	69.6	69.3	69.2	69.0	68.4	68.7	68.0	67.3	66.7	66.4	65.4	64.2	63.6	63.1	63.1	62.3
	R-50	73.0	72.9	72.5	72.1	71.9	71.6	71.6	71.2	71.1	71.0	70.8	69.9	69.1	68.3	68.0	67.8	67.6



(a) ResNet-18 (b) ResNet-50

Figure 5. Comparison with individually trained baselines on CIFAR-100. All scatters are individually trained, whereas our method is trained only once (the red line).



(a) ResNet-18 (b) ResNet-50

Figure 6. Comparison with the original US-Net baseline on CIFAR-100. All are trained only once for 400 epochs.

Table 17. Ablation studies of the loss design under ResNet-18 on CIFAR-100. '-' denotes the model collapses.

Base Loss	Case	Distill Loss	Asymmetric Distill Head	Momentum Target		Linear Accuracy (%)								
				Base model	Sub model	1.0x	0.9x	0.8x	0.7x	0.6x	0.5x	0.4x	0.3x	0.25x
MSE	1	×	×	×	×	-	-	-	-	-	-	-	-	-
	2	MSE	×	×	×	-	-	-	-	-	-	-	-	-
	3	MSE	✓(Share)	×	×	57.5	57.4	57.3	57.0	56.3	55.4	54.5	53.1	52.4
	4	MSE	✓(Share)	×	✓	-	-	-	-	-	-	-	-	-
	5	MSE	×	×	✓	-	-	-	-	-	-	-	-	-
	6	MSE	×	✓	✓	-	-	-	-	-	-	-	-	-
	7	MSE	✓(Share)	✓	✓	64.7	64.7	64.5	64.3	63.9	62.6	61.3	59.7	59.3
	8	MSE	✓(New)	✓	✓	65.4	65.0	64.8	64.5	63.8	62.7	61.1	59.8	58.9
	9	InfoNCE	×	×	×	62.3	62.3	62.3	62.2	61.8	60.6	58.9	57.6	57.2
	10	InfoNCE	✓(Share)	×	×	58.7	58.8	58.8	58.9	58.7	58.4	56.8	55.3	54.3
	11	InfoNCE	✓(New)	×	×	61.5	61.4	61.6	61.6	61.1	60.3	58.7	57.1	56.3
	12	InfoNCE	×	×	✓	63.7	63.8	63.7	63.6	63.1	62.0	60.6	59.3	58.2
	13	InfoNCE	✓(Share)	×	✓	-	-	-	-	-	-	-	-	-
	14	InfoNCE	✓(New)	×	✓	64.5	64.5	64.6	64.5	64.2	63.2	62.1	60.0	59.1
	15	InfoNCE	×	✓	✓	65.0	65.0	65.1	65.0	64.5	62.7	61.3	59.8	59.2
	16	InfoNCE	✓(Share)	✓	✓	65.0	64.9	64.9	64.4	64.1	62.8	61.1	60.0	59.5
	17	InfoNCE	✓(New)	✓	✓	**65.5**	**65.5**	**65.6**	**65.0**	**64.6**	**63.2**	**61.6**	**60.2**	**59.7**
InfoNCE	18	×	×	×	×	64.8	64.0	63.2	62.0	60.8	59.8	57.4	55.1	54.2
	19	MSE	×	×	×	65.0	64.4	63.1	62.3	61.9	60.3	58.3	57.1	56.6
	20	MSE	×	×	✓	65.8	65.0	64.4	63.4	62.7	61.8	59.8	58.5	57.6
	21	MSE	✓	×	✓	66.7	66.0	65.6	64.5	63.3	62.0	60.8	59.3	58.2
	22	MSE	×	✓	✓	66.9	66.3	65.7	64.9	63.8	62.9	61.6	59.5	59.1
	23	MSE	✓	✓	✓	**67.7**	**67.2**	**66.5**	**66.0**	**65.1**	**64.3**	**62.5**	**60.5**	**59.6**
	24	InfoNCE	×	×	×	65.5	64.9	63.8	63.6	62.7	61.8	60.2	58.2	57.4
	25	InfoNCE	×	×	✓	64.7	64.5	64.0	63.6	62.3	61.4	59.8	58.4	57.9
	26	InfoNCE	✓	×	✓	66.1	66.0	65.4	64.4	63.4	62.3	60.8	59.1	58.6
	27	InfoNCE	×	✓	✓	66.0	65.4	64.8	64.3	63.8	62.4	61.1	59.8	58.7
	28	InfoNCE	✓	✓	✓	**67.4**	**66.0**	**66.1**	**65.6**	**64.7**	**64.0**	**62.2**	**60.2**	**59.5**