Three Guidelines You Should Know for Universally Slimmable Self-Supervised Learning

Yun-Hao Cao[1] , Peiqin Sun[2] , Shuchang Zhou[2]

[1] Nanjing University

[2] MEGVII Technology

Poster Session: WED-PM-323

Presented by **Yun-Hao Cao**
2023.04.26

1-Minute Introduction

- Motivation: We aim to train universally slimmable self-supervised networks that can run at arbitrary width to facilitate downstream deployment.

- Challenge: The self-supervised scenario is quite different and directly replacing the supervised loss with self-supervised loss does not work.

Type	Method	Accuracy (%)			
		1.0x	0.75x	0.5x	0.25x
Supervised	Individual	73.8	72.8	71.4	67.3
	S-Net [32]	71.9	71.7	70.8	66.2
	S-Net+Distill [31]	73.1	71.9	70.5	67.2
SimSiam [9]	Individual	65.2	64.0	60.6	51.2
	S-Net [32]	-	-	-	-
	S-Net+Distill [31]	46.9	46.9	46.7	45.3
	Ours	**65.5**	**65.3**	**63.2**	**59.7**

collapse

A big gap

- Our Solution:
  - ➢ We discover that temporal consistent guidance is the key to the success of SSL for universally slimmable networks, and we propose three guidelines for the loss design to ensure this temporal consistency from a unified gradient perspective.
  - ➢ We also propose dynamic sampling and group regularization to simultaneously improve accuracy and training efficiency.

Outline

- Introduction
- Method
- Experimental results
- Conclusions

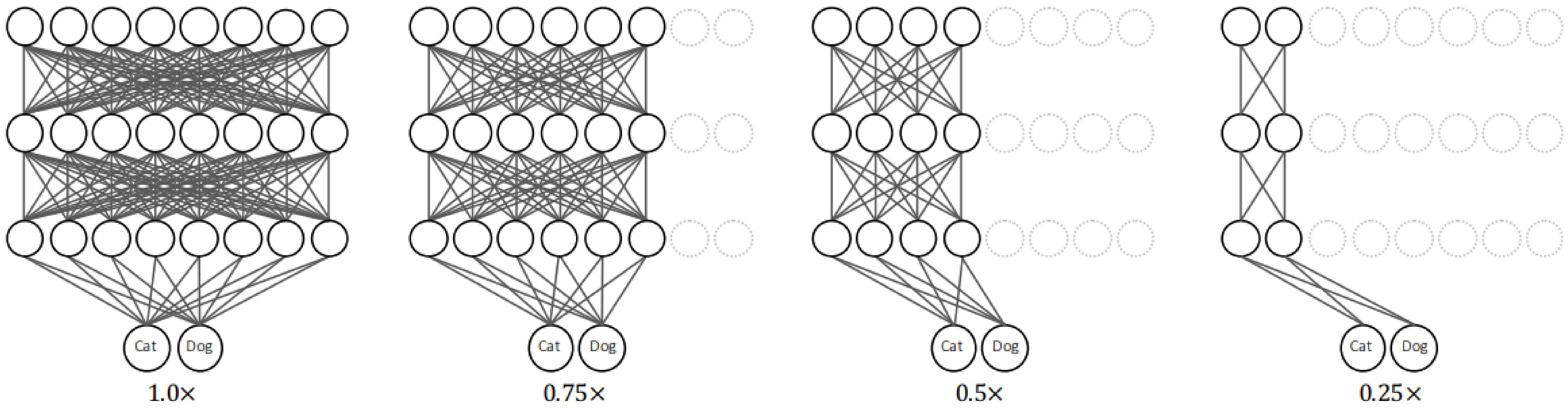
Background and motivation

- With the success of self-supervised learning (SSL), it has become the mainstream paradigm to fine-tune from self-supervised pretrained models to boost the performance on downstream tasks.

Less label dependency!

Better performance!

- (Universally) Slimmable networks can switch freely among different widths by training only once.



- Driven by the success of SSL and slimmable networks, a question arises: Can we train a self-supervised model that can run at arbitrary width?

Background and motivation

- We find that the naive solution (replacing the supervised loss with self-supervised loss) doesn’t work directly after empirical studies.

Table 1. Comparisons between supervised classification and SimSiam under S-Net on CIFAR-100. The accuracy for SimSiam is under linear evaluation. ‘-’ denotes the model collapses.

Type	Method	Accuracy (%)			
		1.0x	0.75x	0.5x	0.25x
Supervised	Individual	73.8	72.8	71.4	67.3
	S-Net [32]	71.9	71.7	70.8	66.2
	S-Net+Distill [31]	73.1	71.9	70.5	67.2
SimSiam [9]	Individual	65.2	64.0	60.6	51.2
	S-Net [32]	-	collapse		-
	S-Net+Distill [31]	46.9	46.9	46.7	45.3
	Ours	**65.5**	**65.3**	**63.2**	**59.7**

A big gap

SimSiam + S-Net: Model collapse

SimSiam + US-Net: Still far from individually trained networks.

Preliminary

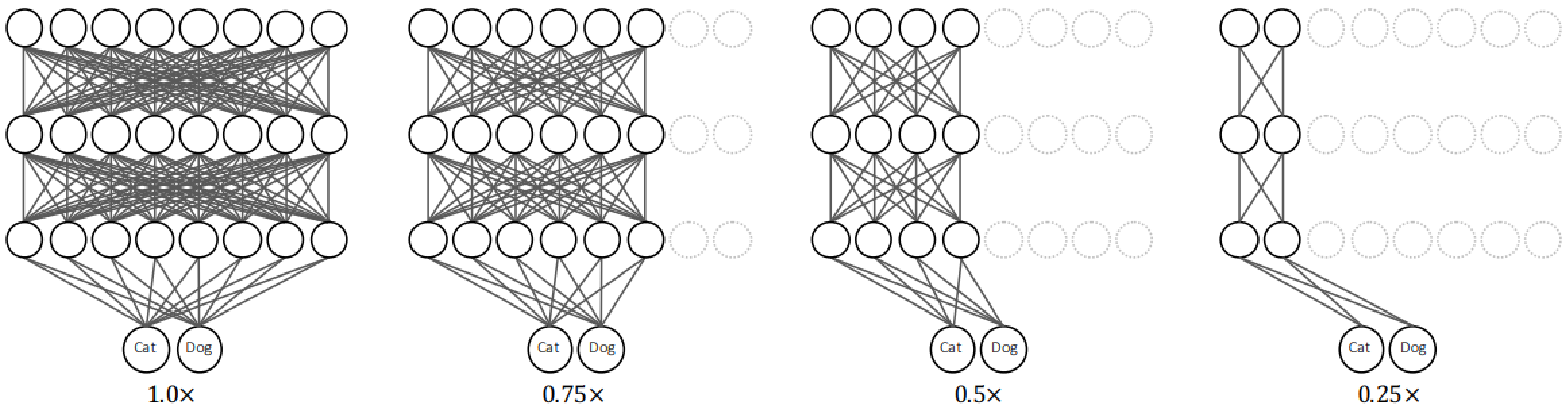
- SimSiam / BYOL: Maximizing similarity between positive samples

$$L_{\text{MSE}} = \sum_i D(\boldsymbol{p}_{i,1}, SG(\boldsymbol{z}_{i,2})) + D(\boldsymbol{p}_{i,2}, SG(\boldsymbol{z}_{i,1}))$$

- SimCLR / MoCo: Contrast with negative samples

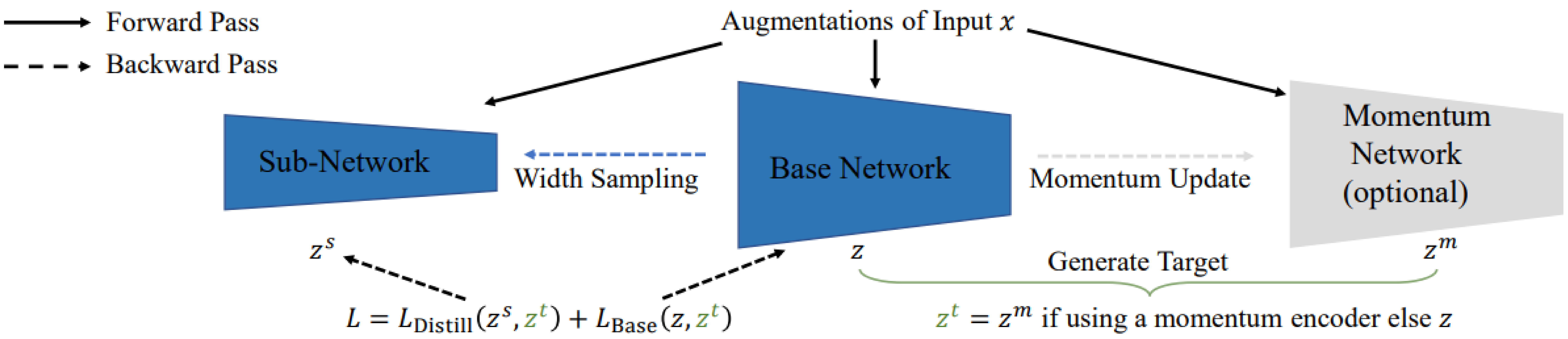
$$L_{\text{NCE}} = -\sum_i \log \frac{e^{\boldsymbol{z}_{i,1}\cdot\boldsymbol{z}_{i,2}}}{e^{\boldsymbol{z}_{i,1}\cdot\boldsymbol{z}_{i,2}} + \sum_{j\neq i, v\in\{1,2\}} e^{\boldsymbol{z}_{i,1}\cdot\boldsymbol{z}_{j,v}}}$$

- (Universally) Slimmable Networks: Base Network Training + Sub-Network Training



US3L: Universally Slimmable Self-Supervised Learning

- Our method

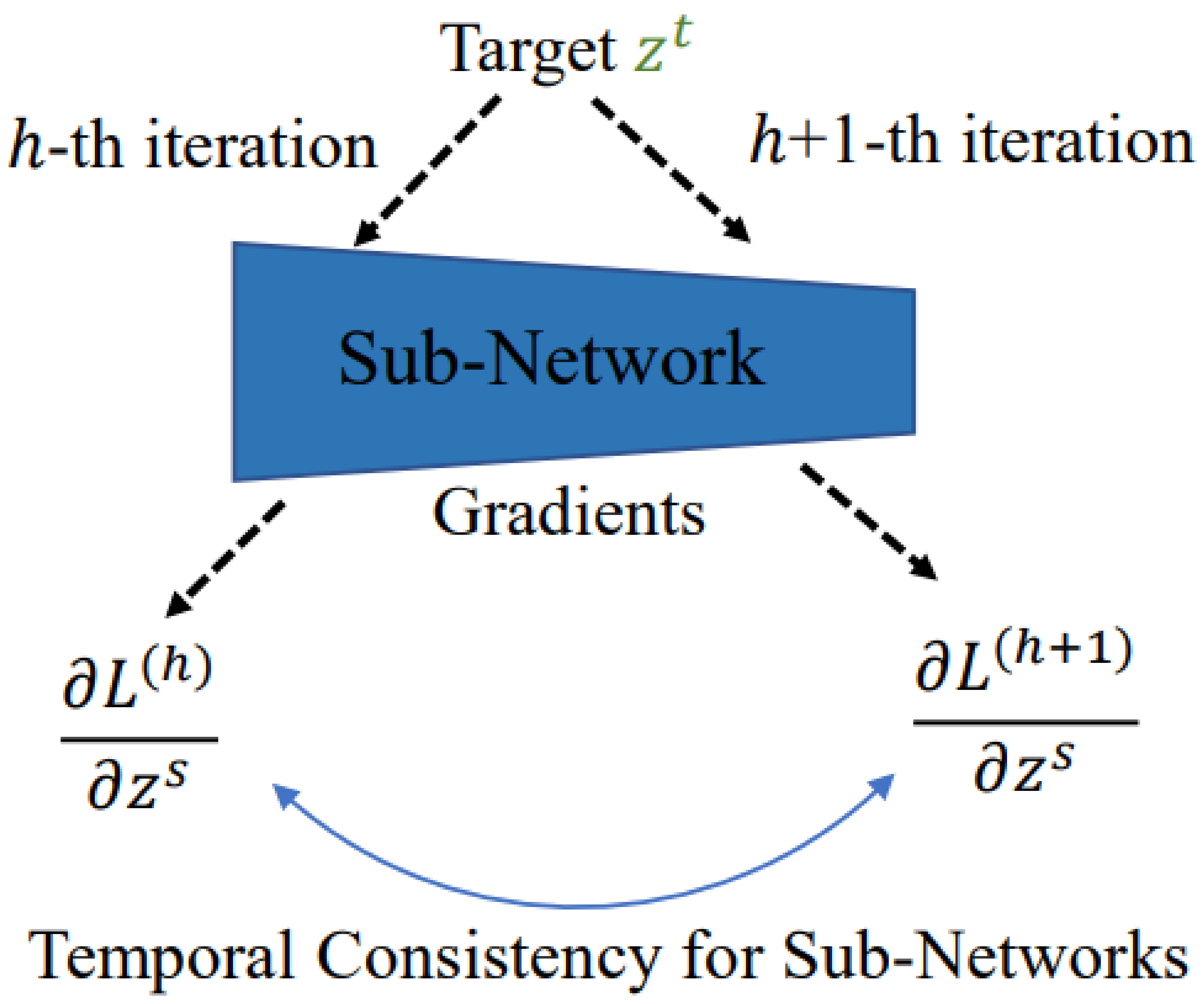


(a) General Framework

$$L = \underbrace{L_{\text{NCE}}}_{L_{\text{Base}}} \underbrace{- \sum_i \sum_s g(z_i^s) \cdot z_i^m}_{L_{\text{Distill}}}$$

Temporal Consistency

- MSE is not robust to changes in the output whereas InfoNCE is stabilized by distances from other samples.

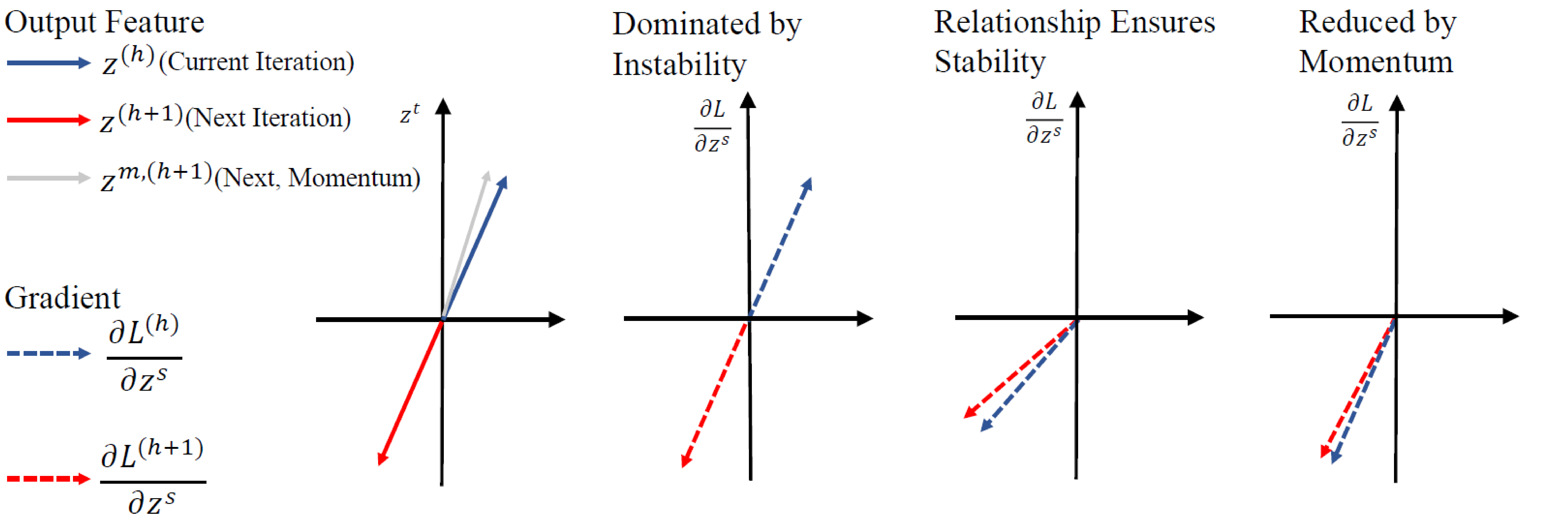


MSE $$\frac{\partial L^{(h+1)}}{\partial z_i^s} - \frac{\partial L^{(h)}}{\partial z_i^s} = (I - w^\theta) z_i^t$$

InfoNCE $$\frac{\partial L^{(h+1)}}{\partial z_i^s} - \frac{\partial L^{(h)}}{\partial z_i^s} = (I - w^\theta)(z_i^t - \sum_j P_j z_j^t)$$

The Proposed Three Guidelines

1. The base loss is based on the relative distance to produce temporal consistent outputs of the base network.
2. The distillation loss is based on the relative distance to produce temporal consistent guidance for sub-networks.
3. A momentum teacher is used to produce stable guidance for sub-networks.

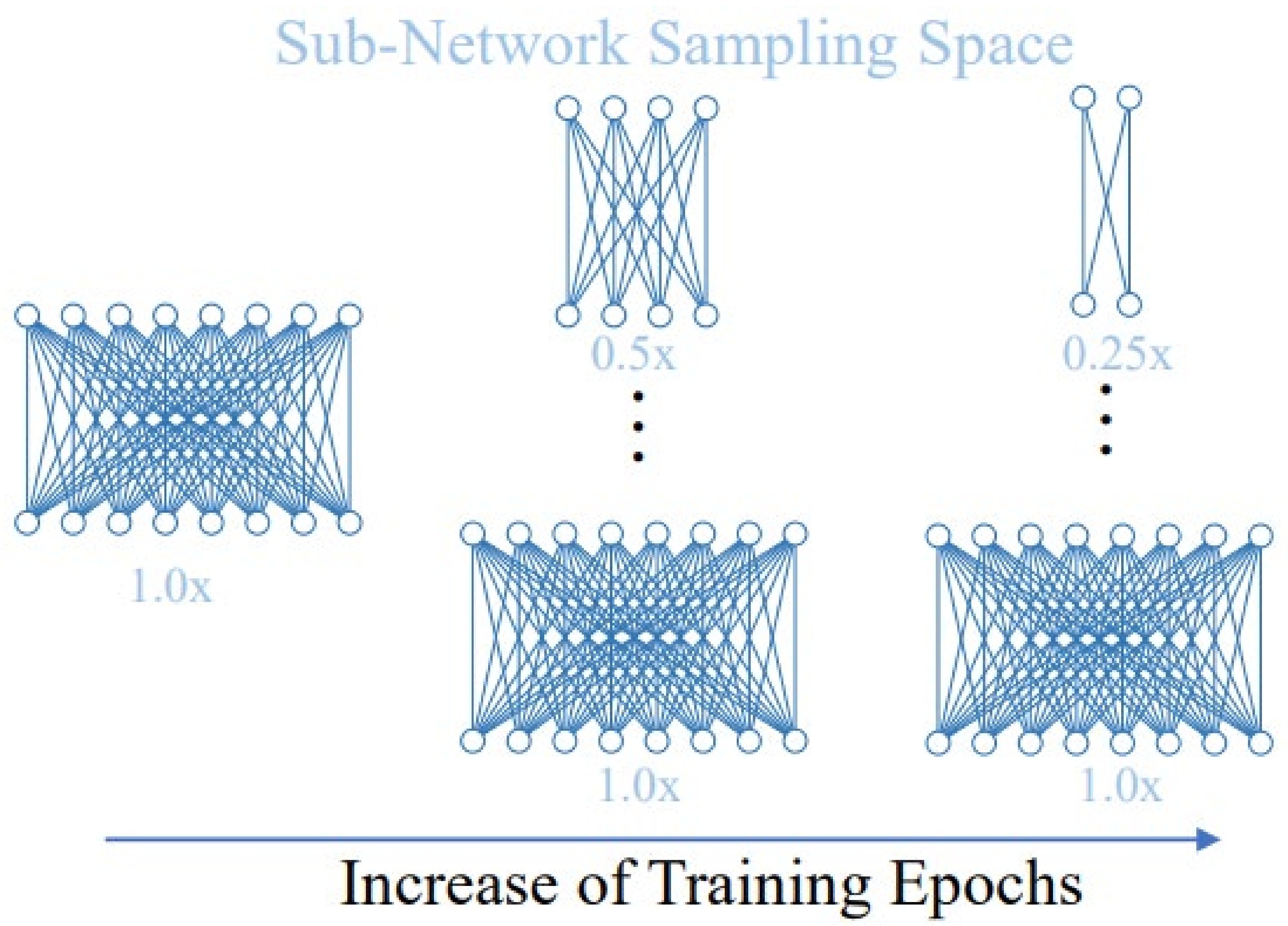


(a) Output Feature (b) MSE (c) InfoNCE (d) MSE+Momentum

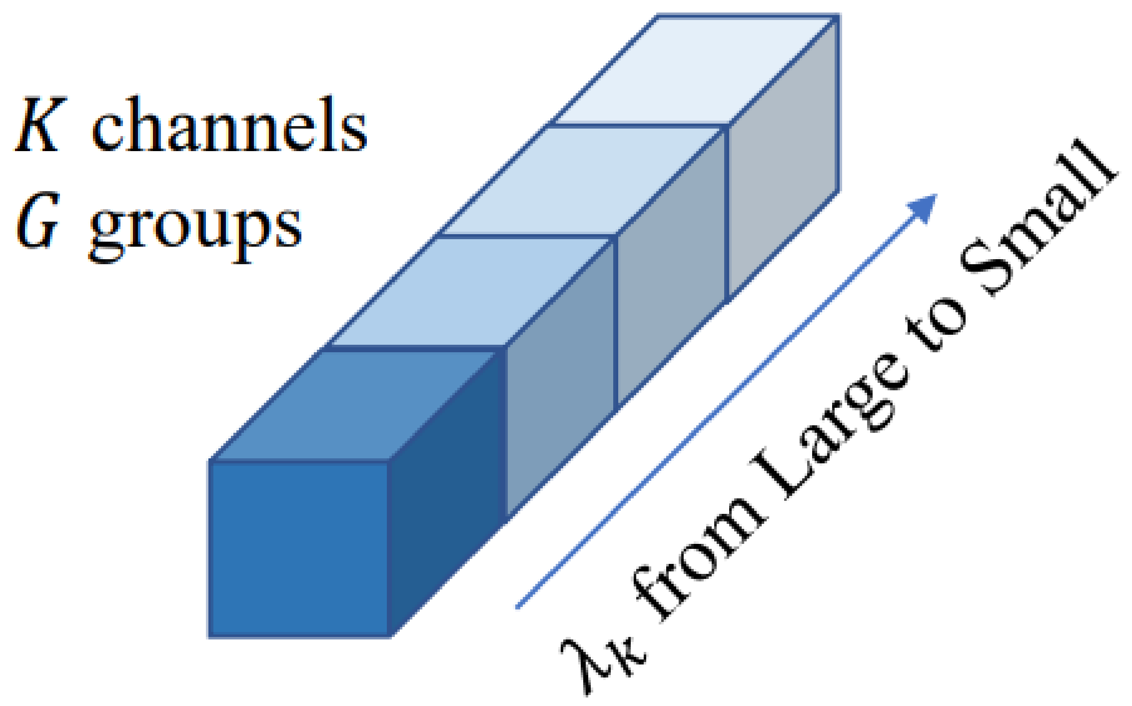
Dynamic Sampling and Group Regularization

- Dynamic Sampling:
  - It is unnecessary to introduce the training of sub-networks at the beginning.
  - The training of sub-networks should be gradual.

[1.0, 1.0] →[0.75, 1.0]→[0.5, 1.0]→[0.25, 1.0]



- Group Regularization:
  - In the training of US-Net, the majority of the weights will be concentrated on the earlier channels.
  - We propose group regularization by giving more degrees of freedom (smaller coefficients) to the later channels.



$$L_{Greg} = \sum_{k=1}^{K} \lambda_k \|w_k\|_2^2$$

Experimental Results

- Our method achieves higher accuracy consistently than baseline methods, with much less training cost.
- Even comparable with SEED, which requires pretrained teacher and individual training.

Backbone	Method	Once Training	Pretrained Teacher	Training Cost	Linear Accuracy (%)								
					1.0x	0.9x	0.8x	0.7x	0.6x	0.5x	0.4x	0.3x	0.25x
ResNet-18	SimCLR [7]	×	×	nT	66.5	65.4	64.7	63.7	62.6	61.0	59.0	56.1	53.6
	SimSiam [9]	×	×	nT	66.5	65.4	64.6	63.5	62.6	60.0	58.3	54.9	52.4
	BYOL [14]	×	×	nT	66.8	66.0	65.6	65.3	63.0	62.1	59.5	56.0	54.3
	SEED [13]	×	BYOL R-50	nT	67.3	66.6	65.8	65.2	64.8	63.5	62.2	60.1	58.5
	SEED-MSE	×	BYOL R-18	nT	67.5	67.2	66.5	66.0	65.9	64.8	64.0	62.4	60.1
	SEED-MSE	×	BYOL R-50	nT	67.5	66.8	66.7	66.0	65.4	64.9	63.6	61.3	60.1
	US [31]+SimCLR	✓	×	$4T$	65.5	64.9	63.8	63.6	62.7	61.8	60.2	58.2	57.4
	US [31]+SimSiam	✓	×	$4T$	57.5	57.4	57.3	57.0	56.3	55.4	54.5	53.1	52.4
	Ours	✓	×	**$2.5T$**	69.0	68.2	68.0	66.9	66.1	64.7	62.6	60.9	60.4
	Ours (800ep)	✓	×	$5T$	**70.1**	**69.3**	**69.0**	**68.7**	**67.3**	**66.4**	**64.2**	**63.1**	**62.3**
ResNet-50	BYOL [14]	×	×	nT	67.0	66.7	66.5	66.3	66.0	64.9	63.8	62.1	61.2
	SEED [13]	×	BYOL R-50	nT	70.3	69.8	69.6	69.4	69.0	68.2	67.2	65.6	65.1
	SEED-MSE	×	BYOL R-50	nT	69.4	69.0	68.5	69.1	68.4	68.1	67.3	66.9	66.4
	US [31]+SimCLR	✓	×	$4T$	70.1	69.9	69.7	69.3	68.7	68.2	67.5	66.0	65.5
	US [31]+SimSiam	✓	×	$4T$	54.7	54.6	54.7	54.7	54.7	54.8	54.6	54.3	54.0
	Ours	✓	×	**$2.5T$**	72.6	72.0	71.5	71.2	70.6	70.2	68.6	67.7	67.4
	Ours (800ep)	✓	×	$5T$	**73.0**	**72.5**	**71.9**	**71.6**	**71.1**	**70.8**	**69.1**	**68.0**	**67.6**
MobileNetv2	BYOL [14]	×	×	nT	61.2	60.7	60.5	60.2	59.9	58.7	57.3	54.6	51.9
	SEED-MSE	×	BYOL R-50	nT	**68.6**	**68.9**	**67.6**	**67.3**	**67.4**	**66.3**	**65.5**	**64.0**	**62.6**
	SEED-MSE	×	BYOL MBv2	nT	63.8	63.5	63.8	63.6	63.6	63.3	62.7	62.1	59.8
	US [31]+SimCLR	✓	×	$4T$	56.2	56.0	55.3	55.0	54.8	54.3	54.0	53.2	52.2
	US [31]+SimSiam	✓	×	$4T$	-	-	-	-	-	-	-	-	-
	Ours	✓	×	**$2.5T$**	65.7	65.1	64.2	63.6	63.4	62.2	61.5	60.7	59.3

Application to Vision Transformers

- Effectiveness when applied to vision transformer? Yes.

Table 3. Linear evaluation results for ViT on CIFAR-10.

Backbone	Method	Once Training	Linear Accuracy (%)			
			1.0x	0.75x	0.5x	0.25x
ViT-Tiny	MoCov3 [10]	×	82.6	79.5	75.8	68.0
	US+MoCov3	✓	79.8	79.4	77.6	76.4
	Ours	✓	**86.0**	**84.7**	**83.3**	**80.2**
ViT-Small	MoCov3 [10]	×	88.0	86.8	83.0	75.5
	US+MoCov3	✓	88.2	87.5	86.3	84.9
	Ours	✓	**90.3**	**89.7**	**88.7**	**85.5**

ImageNet and Transferring Experiments

- US3L achieves better performance at all widths with only once training and one copy of weights.

Table 4. Linear evaluation results on ImageNet.

Backbone	Method	Once Training	Linear Accuracy (%) 1.0x	0.75x	0.5x	0.25x
ResNet-18	BYOL	×	54.0	53.7	47.4	34.9
	US+BYOL	✓	55.9	53.1	48.0	40.6
	Ours	✓	**56.9**	**54.5**	**48.7**	**40.7**
ResNet-50	BYOL	×	68.1	66.3	61.2	50.9
	US+BYOL	✓	64.7	64.3	62.6	57.1
	Ours	✓	**68.4**	**66.7**	**63.4**	**57.7**

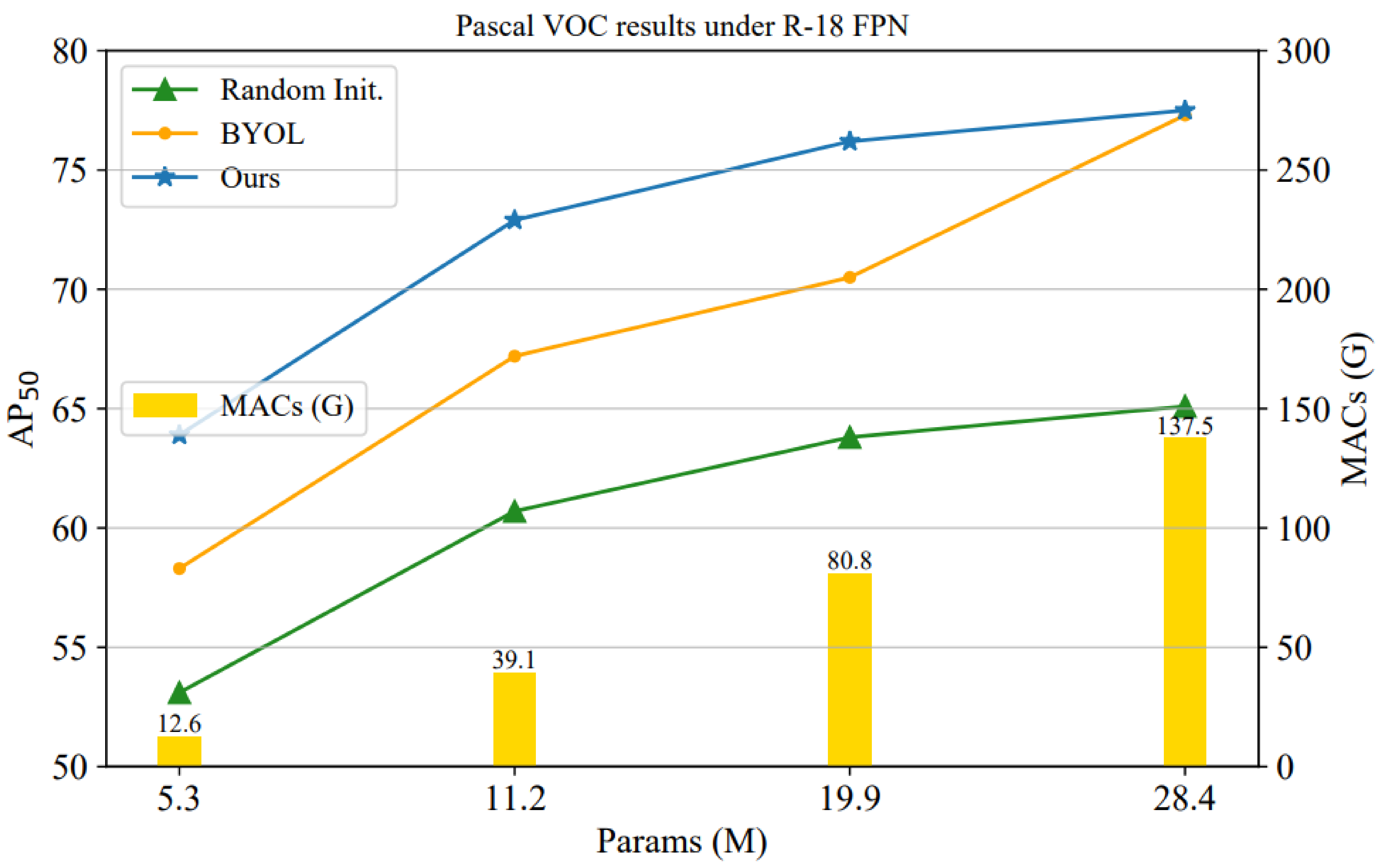
- Similar trends are observed when transferring to downstream classification tasks.

Table 6. Transfer results on recognition benchmarks under linear evaluation. 'C-10/100' denotes 'CIFAR-10/100'.

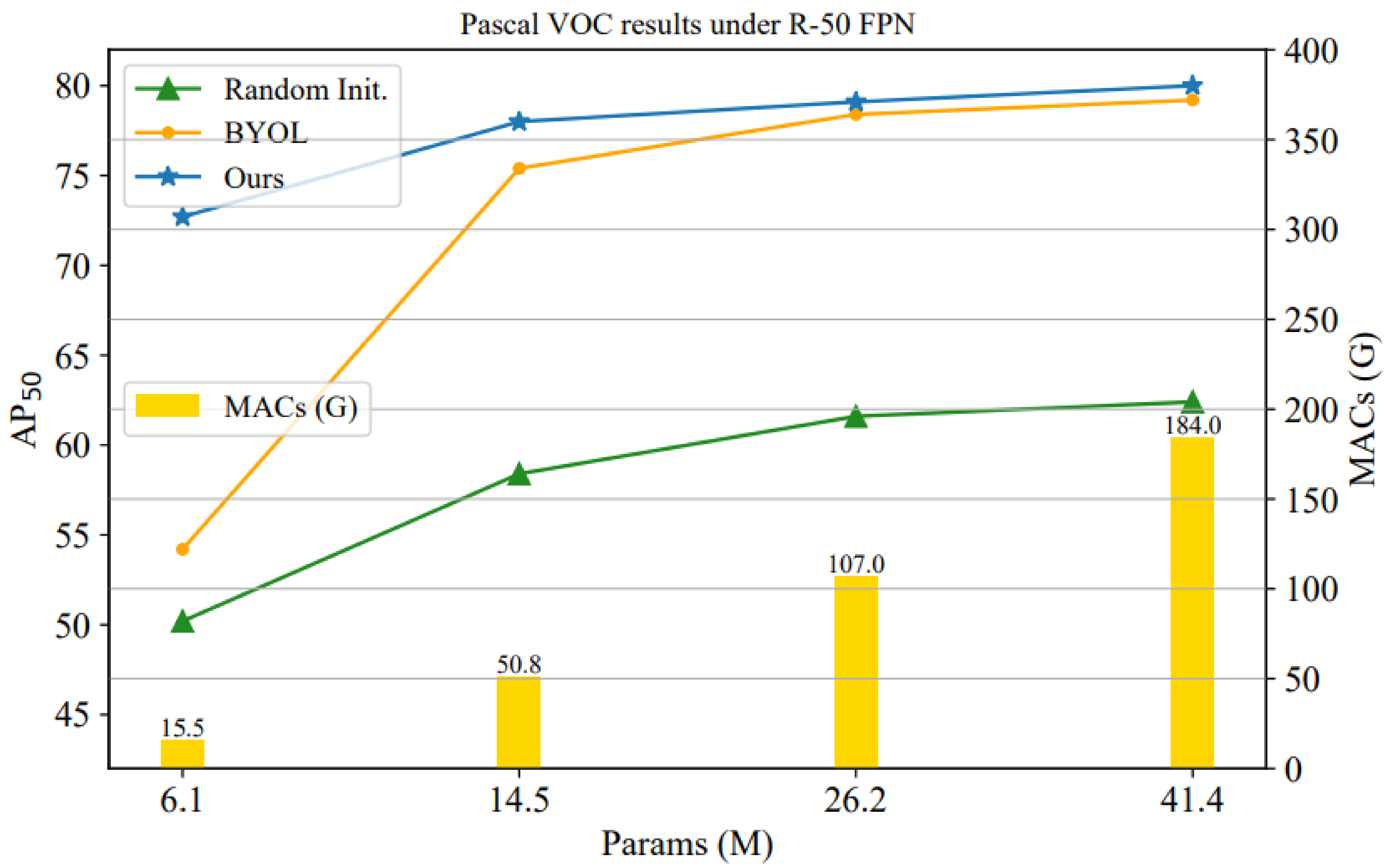
Net	Width	Params	MACS	Method	Linear Accuracy (%) C-10	C-100	Flowers	Pets	Dtd
R-50	1.0x	22.56M	4.11G	BYOL	**87.1**	60.6	81.0	**80.9**	70.7
				Ours	**87.1**	**61.5**	**90.6**	79.4	**72.6**
	0.75x	14.77M	2.34G	BYOL	83.6	52.8	82.9	74.2	66.3
				Ours	**84.4**	**56.9**	**89.7**	**78.0**	**71.1**
	0.5x	6.92M	1.06G	BYOL	80.6	52.0	74.8	75.0	65.9
				Ours	**81.6**	**52.8**	**88.1**	**76.8**	**68.8**
	0.25x	1.99M	0.28G	BYOL	75.9	46.2	75.4	64.7	61.4
				Ours	**78.9**	**49.9**	**84.4**	**74.0**	**64.9**

Downstream Object Detection

- As we decrease the width, our advantages over the baseline counterpart BYOL will be further expanded.



R-18 FPN



R-50 FPN

Ablation studies

- Experimental results are in full agreement with the proposed three guidelines.
- Consistency should not only exist between iterations, but also across sub-networks.
- The use of an auxiliary distillation head will result in consistent improvements.
- ……

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

Ablation studies

- Dynamic sampling and group regularization both improves the accuracy for various backbones.

Backbone	Dynamic Sampling	Group Reg.	Linear Accuracy (%) 1.0x	0.8x	0.6x	0.5x	0.3x	0.25x
R-18	×	×	67.7	66.5	65.1	64.3	60.5	59.6
	✓	×	68.6	67.2	65.5	64.6	60.7	59.9
	×	✓	68.6	67.3	65.5	64.4	**60.9**	60.1
	✓	✓	**69.0**	**68.0**	**66.1**	**64.7**	**60.9**	**60.4**
R-50	×	×	71.0	70.6	70.0	69.1	67.2	66.8
	✓	×	71.8	71.1	70.2	69.3	67.3	67.2
	×	✓	71.9	71.1	70.0	69.6	**67.7**	**67.5**
	✓	✓	**72.6**	**71.5**	**70.6**	**70.2**	**67.7**	67.4
MBv2	×	×	62.9	62.0	61.5	60.4	59.6	58.7
	✓	×	64.7	63.3	62.3	61.7	**60.7**	59.2
	×	✓	64.0	63.2	62.1	61.4	60.2	59.0
	✓	✓	**65.7**	**64.2**	**63.4**	**62.2**	**60.7**	**59.3**

- Our dynamic sampling strategy can be used alone or combined with the sandwich rule.

Sandwich Rule	Dynamic Sampling	Linear Accuracy (%) 1.0x	0.8x	0.6x	0.5x	0.3x	0.25x
×	×	65.1	65.0	64.7	63.2	59.4	56.4
×	✓	67.4	**67.3**	**65.9**	**64.7**	59.9	58.7
✓	×	67.7	66.5	65.1	64.3	60.5	59.6
✓	✓	**68.6**	**67.3**	65.5	64.4	**60.9**	**60.1**

Conclusions

- ✓ We discovered significant differences between supervised and self-supervised learning when training US-Net. Based on these observations, we analyzed and summarized three guidelines for the loss design.
- ✓ We proposed a dynamic sampling strategy to reduce the training cost without sacrificing accuracy, which eases coping with the large data volumes in SSL.
- ✓ We analyzed how the training scheme of US-Net limits the model capacity and proposed group regularization.
- ✓ Exhaustive experimental results further show that our US3L achieves better performance on various benchmarks at all widths.

Thanks!

Q&A