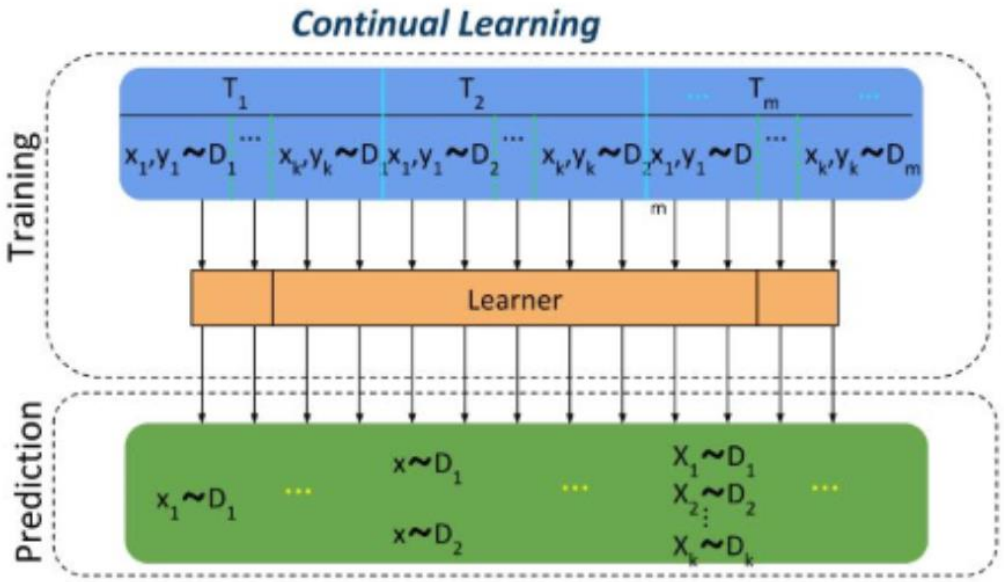
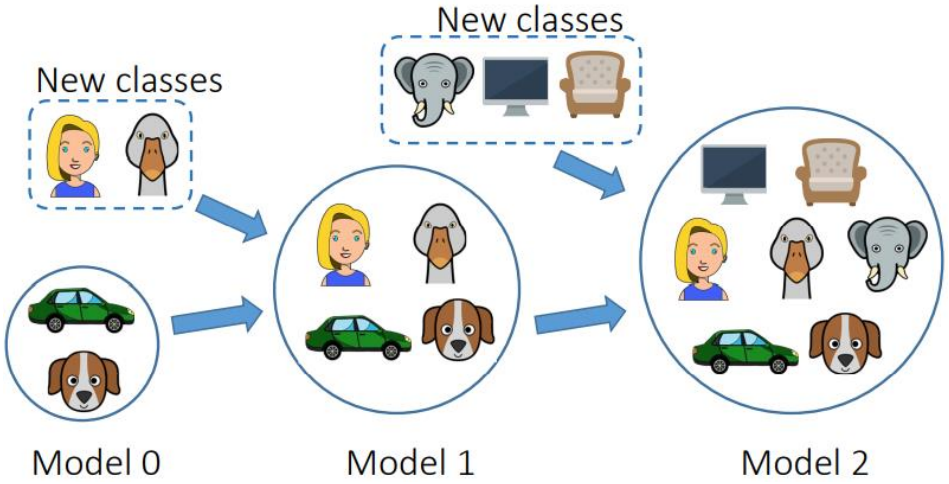
Meta-attention for ViT-backed Continual Learning

Mengqi Xue[1], Haofei Zhang[1], Jie Song[1,†], Mingli Song[1,2]
[1]Zhejiang University
[2]Alibaba-Zhejiang University Joint Research Institute of Frontier Technologies, Zhejiang University
{mqxue,haofeizhang,sjie,brooksong}@zju.edu.cn

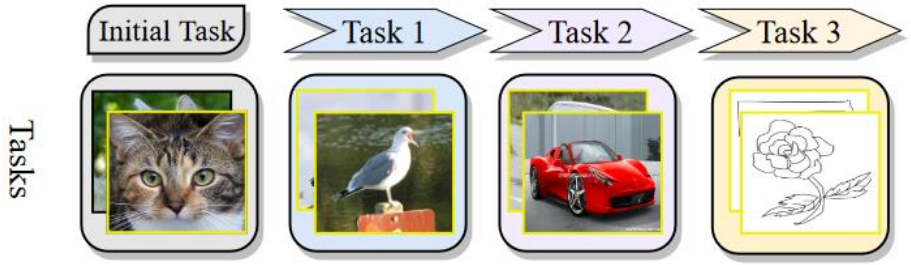
Incremental learning



Class Incremental learning



Task Incremental learning

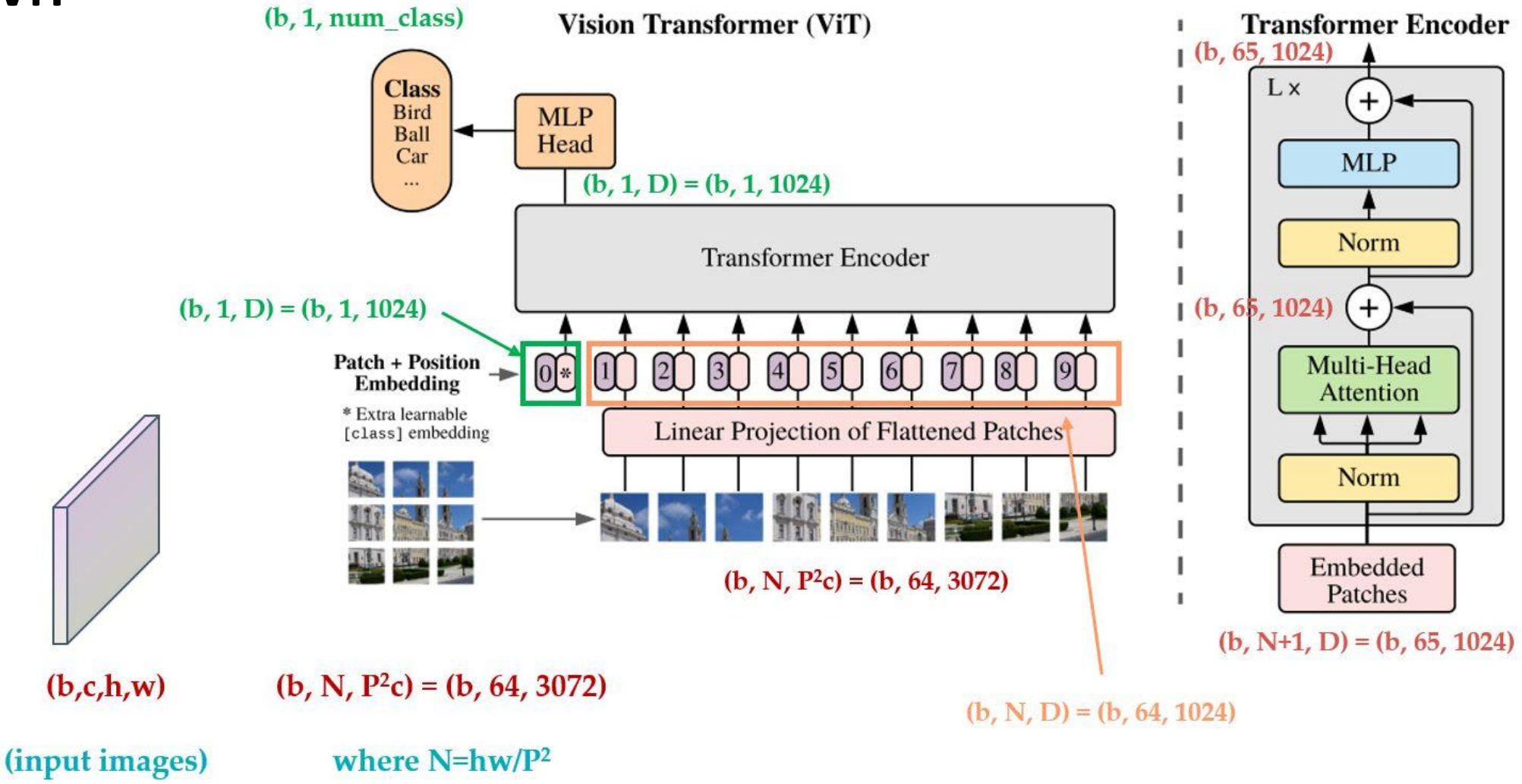


Incremental learning (Continual learning , lifelong learning):

- independent and identically distributed(独立同分布, i.i.d.)
- Stability-plasticity Dilemma

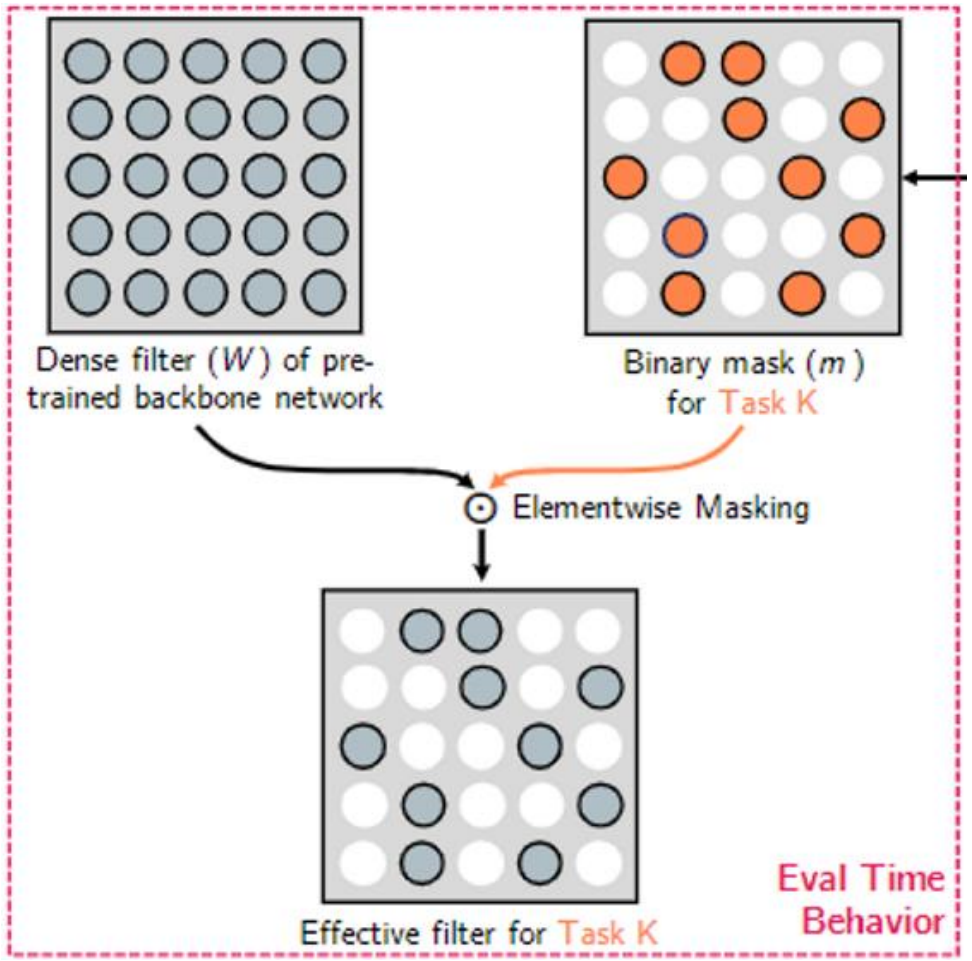
· Learning Step
· predictions with all seen categories/tasks

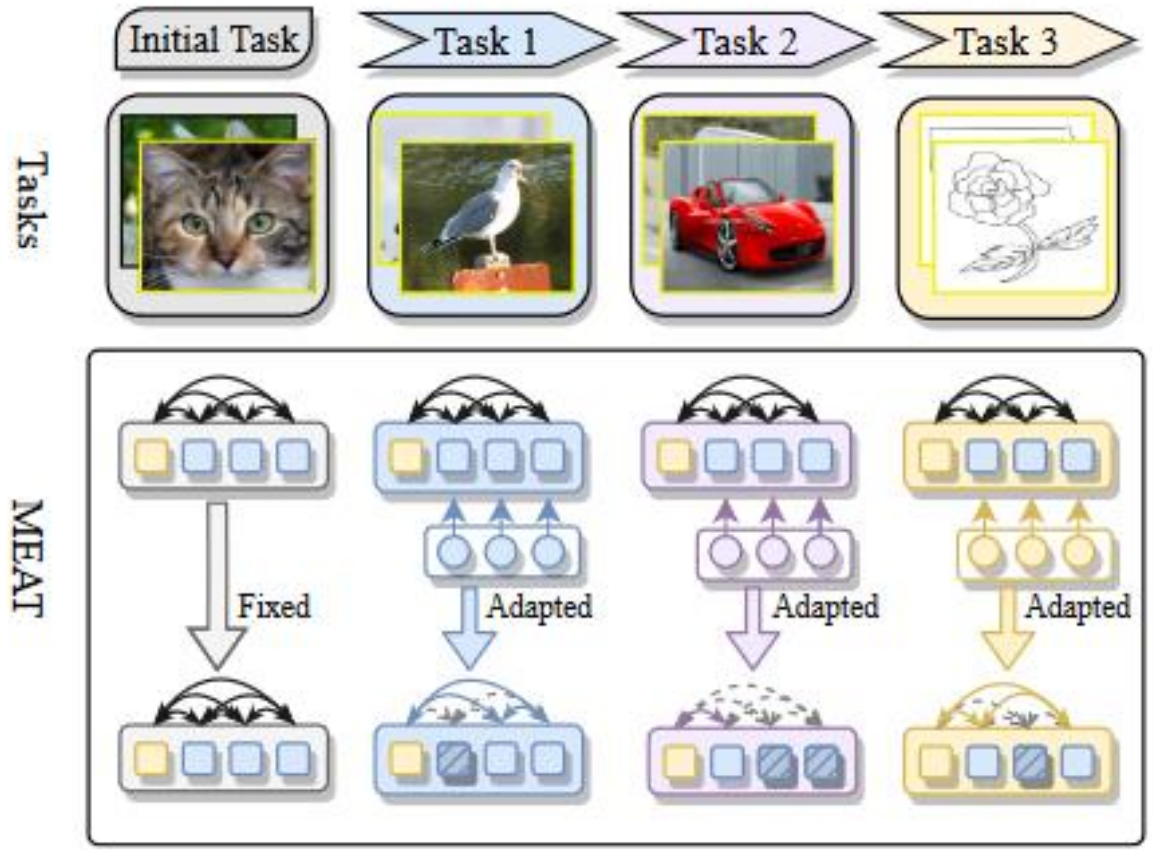
ViT



Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[J]. arXiv preprint arXiv:2010.11929, 2020.

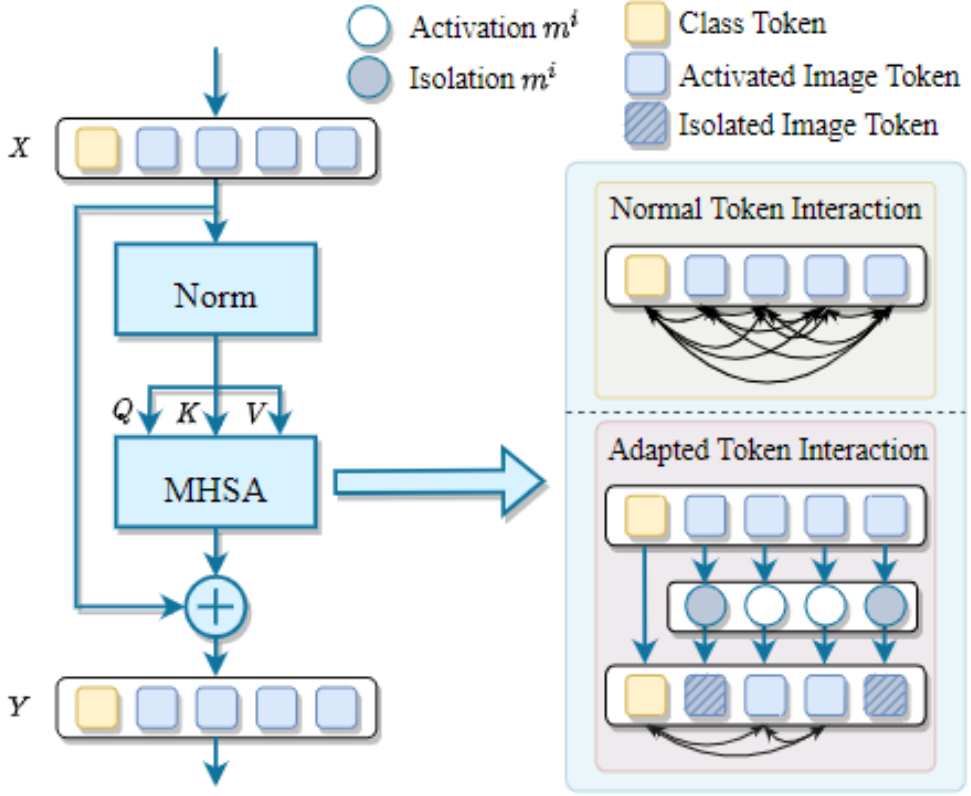
Mask methods





Mallya A, Davis D, Lazebnik S. Piggyback: Adapting a single network to multiple tasks by learning to mask weights[C]//Proceedings of the European Conference on Computer Vision (ECCV). 2018: 67-82.

Meta attention



Attention to attention

Generate **binary** mask to control self-Attention:

Learnable parameters $t \in \mathbb{R}^2$

Gumbel-Softmax estimator:

$$m^i = \frac{\exp((\log(t^{i,1}) + g^1)/\tau)}{\sum_{k=1}^{2} \exp((\log(t^{i,k}) + g^k)\tau)},$$

Objective

A. $\mathcal{L}_{ce}$

B. $\mathcal{L}_{dc}(m) = \frac{1}{L}\sum_{l=1}^{L}\left(\lambda - \frac{1}{n}\sum_{i=1}^{n} m_l^i\right)^2,$

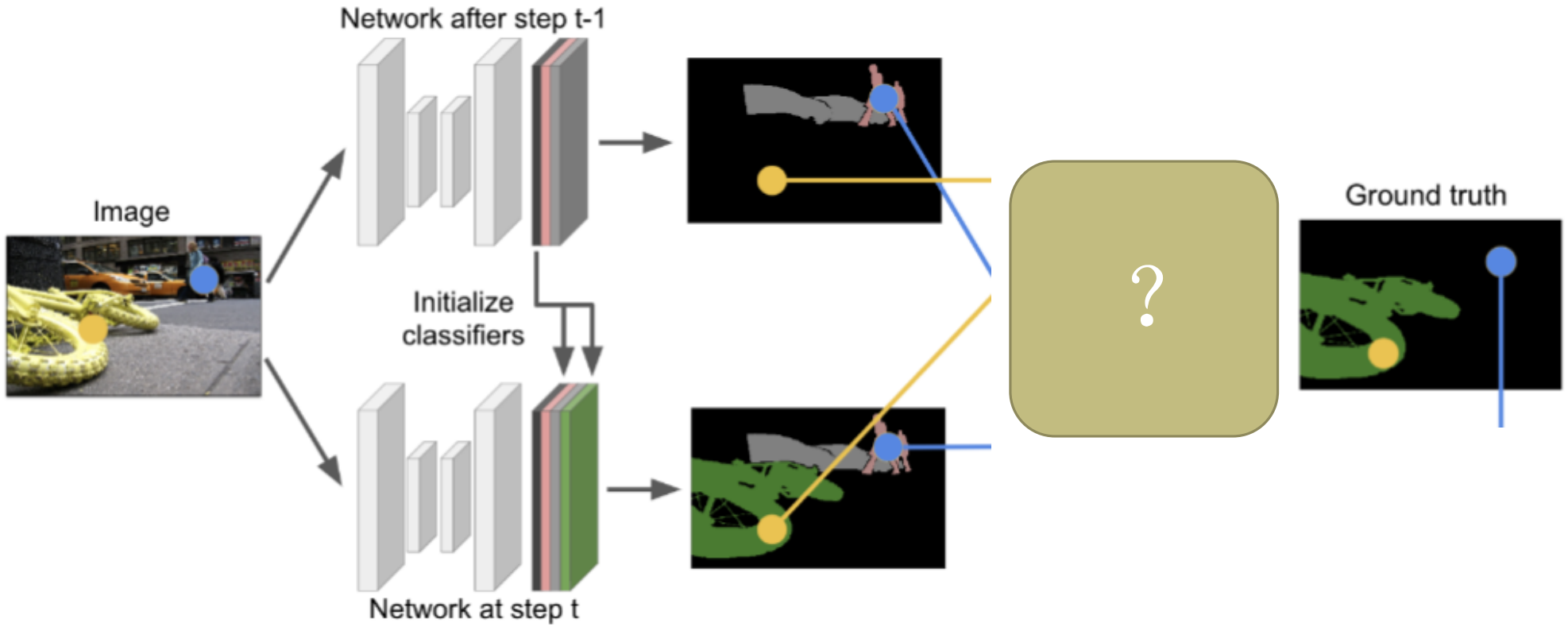
Drop-control loss is to prevent excessive patch dropping with hyper-parameter λ=0.9

	Dataset	Method						
		Individual	Classifier	LwF [27]	Piggyback [30]	HAT [41]	Adaptor-B [16]	MEAT
DeiT-Ti	CUB	75.13	46.05	59.03	60.65	68.34	66.03	**71.16**
	Cars	69.82	16.27	39.39	44.87	50.57	45.50	**53.42**
	FGVC	70.00	14.35	38.87	45.58	46.71	41.28	**52.69**
	WikiArt	72.13	38.64	46.88	62.42	61.84	57.04	**64.63**
	Sketches	73.50	30.64	53.17	69.07	65.49	69.21	**70.73**
	CIFAR-100	83.85	66.05	69.79	71.18	70.67	75.21	**78.13**
	ImageNet	30.82 (0.00)	72.20 (0.00)	26.24 (↓ 45.96)	72.20 (0.00)	N/A N/A	72.20 (0.00)	72.20 (0.00)
	Model Size	149 MB (6.49x)	23 MB (0.06x)	23 MB (1.00x)	26 MB (0.21x)	23 MB (1.01x)	29 MB (0.28x)	25 MB (0.16x)
DeiT-S	CUB	82.69	49.10	69.34	72.89	79.67	77.20	**81.53**
	Cars	84.74	18.29	74.00	74.72	73.22	67.23	**77.20**
	FGVC	82.69	15.51	55.99	60.04	62.99	57.04	**65.69**
	WikiArt	79.48	43.85	65.64	68.09	70.43	71.33	**73.43**
	Sketches	80.68	39.80	70.74	75.03	74.97	72.87	**76.68**
	CIFAR-100	89.03	72.71	75.67	79.76	79.52	84.00	**85.93**
	ImageNet	49.78 (0.00)	79.84 (0.00)	23.01 (↓ 56.83)	79.84 (0.00)	N/A N/A	79.84 (0.00)	79.84 (0.00)
	Model Size	582 MB (6.77x)	86 MB (0.03x)	86 MB (1.00x)	101 MB (0.15x)	86 MB (1.01x)	99 MB (0.17x)	96 MB (0.14x)
T2T-ViT-12	CUB	74.47	26.15	45.33	63.57	66.57	64.31	**69.90**
	Cars	72.67	11.52	59.01	58.22	54.63	53.79	**61.90**
	FGVC	64.09	12.46	42.07	51.47	52.69	48.02	**53.55**
	WikiArt	73.51	35.57	51.24	60.34	58.53	59.01	**61.20**
	Sketches	76.60	18.79	61.98	73.07	71.29	74.02	**74.75**
	CIFAR-100	85.03	33.10	66.34	70.98	74.86	73.58	**77.42**
	ImageNet	32.62 (0.00)	55.42 (0.00)	28.54 (↓ 26.88)	55.42 (0.00)	N/A N/A	55.42 (0.00)	55.42 (0.00)
	Model Size	179 MB (6.63x)	27 MB (0.07x)	27 MB (1.00x)	32 MB (0.20x)	28 MB (1.02x)	36 MB (0.30x)	30 MB (0.14x)

Representation Compensation Networks for Continual Semantic Segmentation

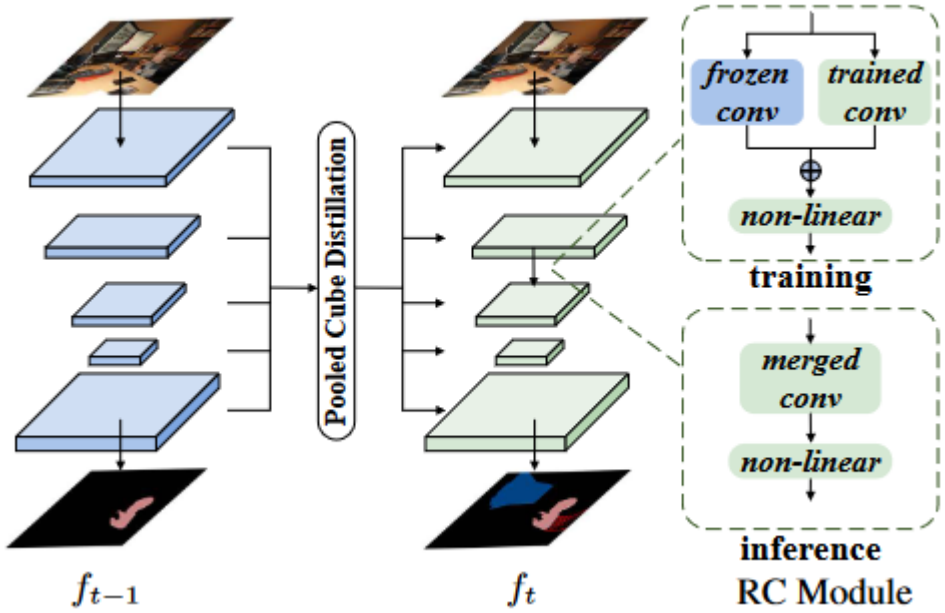
Chang-Bin Zhang [1*] Jia-Wen Xiao [1*] Xialei Liu [1†] Ying-Cong Chen [2] Ming-Ming Cheng [1]
[1] TKLNDST, CS, Nankai University [2] The Hongkong University of Science and Technology

Overview



Cermelli F, Mancini M, Bulo S R, et al. Modeling the background for incremental learning in semantic segmentation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 9233-9242.

Pooled Cube Distillation



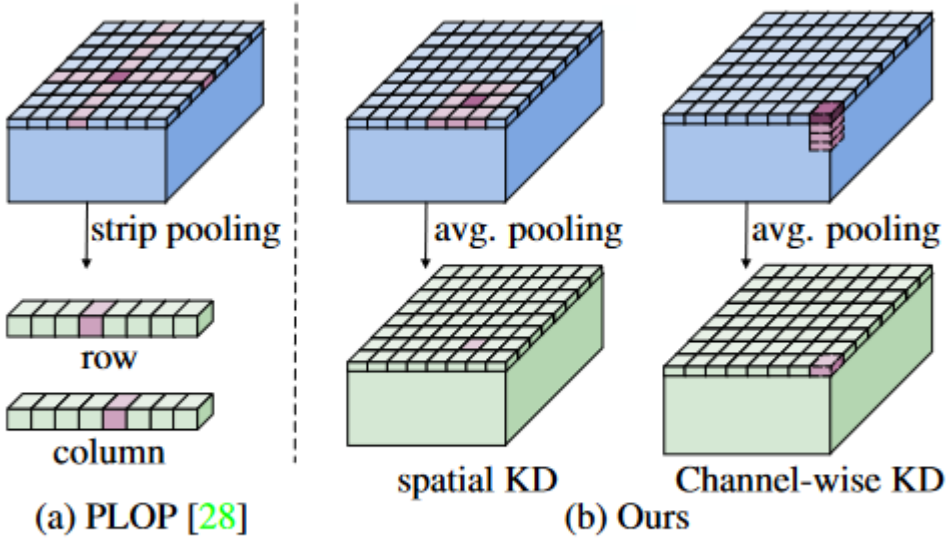
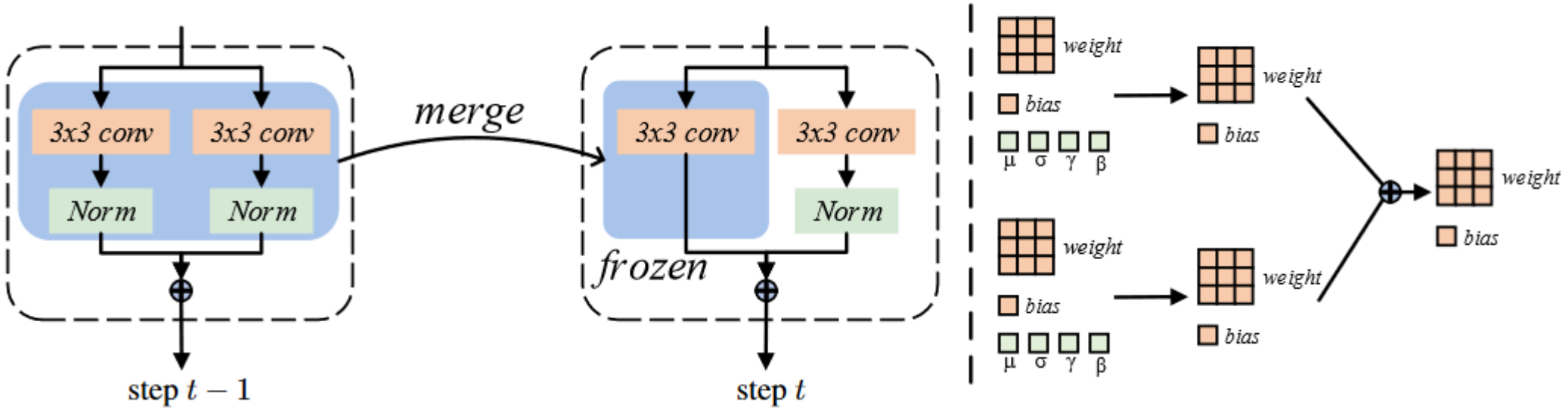


Figure 4. Comparison between PLOP [28] and our proposed Pooled Cube Knowledge Distillation mechanism.

MiB: per-pixel knowledge distillation (KD)
PLOP: row & column KD & Local POD
RCIL: multi-scale & spatial KD & channel KD

RC Module



$$\hat{x} = \sum_{i=0}^{1} Norm_i(W_i x + b_i)$$

$$= \sum_{i=0}^{1} (\gamma_i \frac{W_i x + b_i - \mu_i}{\sigma_i} + \beta_i)$$

$$= (\sum_{i=0}^{1} \frac{\gamma_i W_i}{\sigma_i}) x + \sum_{i=0}^{1} (\frac{\gamma_i b_i - \gamma_i \mu_i}{\sigma_i} + \beta_i)$$

$$= \hat{W} x + \hat{b}.$$

Training:

$$\hat{x} = \eta \cdot x_1 + (1 - \eta) \cdot x_2,$$

$$\eta \in \{0, 0.5, 1\},$$

Inference:

$$\eta = 0.5,$$

Result

Method	15-5 (2 steps)						15-1 (6 steps)						10-1 (11 steps)					
	Disjoint			Overlapped			Disjoint			Overlapped			Disjoint			Overlapped		
	0-15	*16-20*	*all*	*0-15*	*16-20*	*all*	*0-15*	*16-20*	*all*	*0-15*	*16-20*	*all*	*0-10*	*11-20*	*all*	*0-10*	*11-20*	*all*
Fine-tuning	5.7	33.6	12.3	6.6	33.1	12.9	4.6	1.8	3.8	4.6	1.8	3.9	6.3	1.1	3.8	6.4	1.2	3.9
Joint	78.2	78.0	78.2	78.2	78.0	78.2	79.8	72.6	78.2	79.8	72.6	78.2	79.8	72.6	78.2	79.8	72.6	78.2
LwF [52]	60.4	37.4	54.9	60.8	36.6	55.0	5.8	3.6	5.3	6.0	3.9	5.5	7.2	1.2	4.3	8.0	2.0	4.8
ILT [63]	64.9	39.5	58.9	67.8	40.6	61.3	8.6	5.7	7.9	9.6	7.8	9.2	7.3	3.2	5.4	7.2	3.7	5.5
MiB [9]	73.0	43.3	65.9	76.4	49.4	70.0	48.4	12.9	39.9	38.0	13.5	32.2	9.5	4.1	6.9	20.0	20.1	20.1
SDR [64]	74.6	44.1	**67.3**	76.3	50.2	70.1	59.4	14.3	**48.7**	47.3	14.7	39.5	17.3	11.0	**14.3**	32.4	17.1	25.1
PLOP [28]	71.0	42.8	64.3	75.7	51.7	**70.1**	57.9	13.7	46.5	65.1	21.1	**54.6**	9.7	7.0	8.4	44.0	15.5	**30.5**
Ours	75.0	42.8	**67.3**	78.8	52.0	**72.4**	66.1	18.2	**54.7**	70.6	23.7	**59.4**	30.6	4.7	**18.2**	55.4	15.1	**34.3**

Table 1. The mIoU(%) of the last step on the Pascal VOC 2012 dataset for different continual class segmentation scenarios. The **red** denotes the highest results and the **blue** denotes the second highest results.

Method	100-50 (2 steps)			100-10 (6 steps)							50-50 (3 steps)			
	1-100	*101-150*	*all*	*1-100*	*101-110*	*111-120*	*121-130*	*131-140*	*141-150*	*all*	*1-50*	*51-100*	*101-150*	*all*
ILT [63]	18.3	14.8	17.0	0.1	0.0	0.1	0.9	4.1	9.3	1.1	13.6	12.3	0.0	9.7
MiB [9]	40.7	17.7	32.8	38.3	12.6	10.6	8.7	9.5	15.1	29.2	45.3	26.1	17.1	29.3
PLOP [28]	41.9	14.9	**32.9**	40.6	15.2	16.9	18.7	11.9	7.9	**31.6**	48.6	30.0	13.1	**30.4**
Ours	42.3	18.8	**34.5**	39.3	14.6	26.3	23.2	12.1	11.8	**32.1**	48.3	31.3	18.7	**32.5**
Joint	44.3	28.2	38.9	44.3	26.1	42.8	26.7	28.1	17.3	38.9	51.1	38.3	28.2	38.9

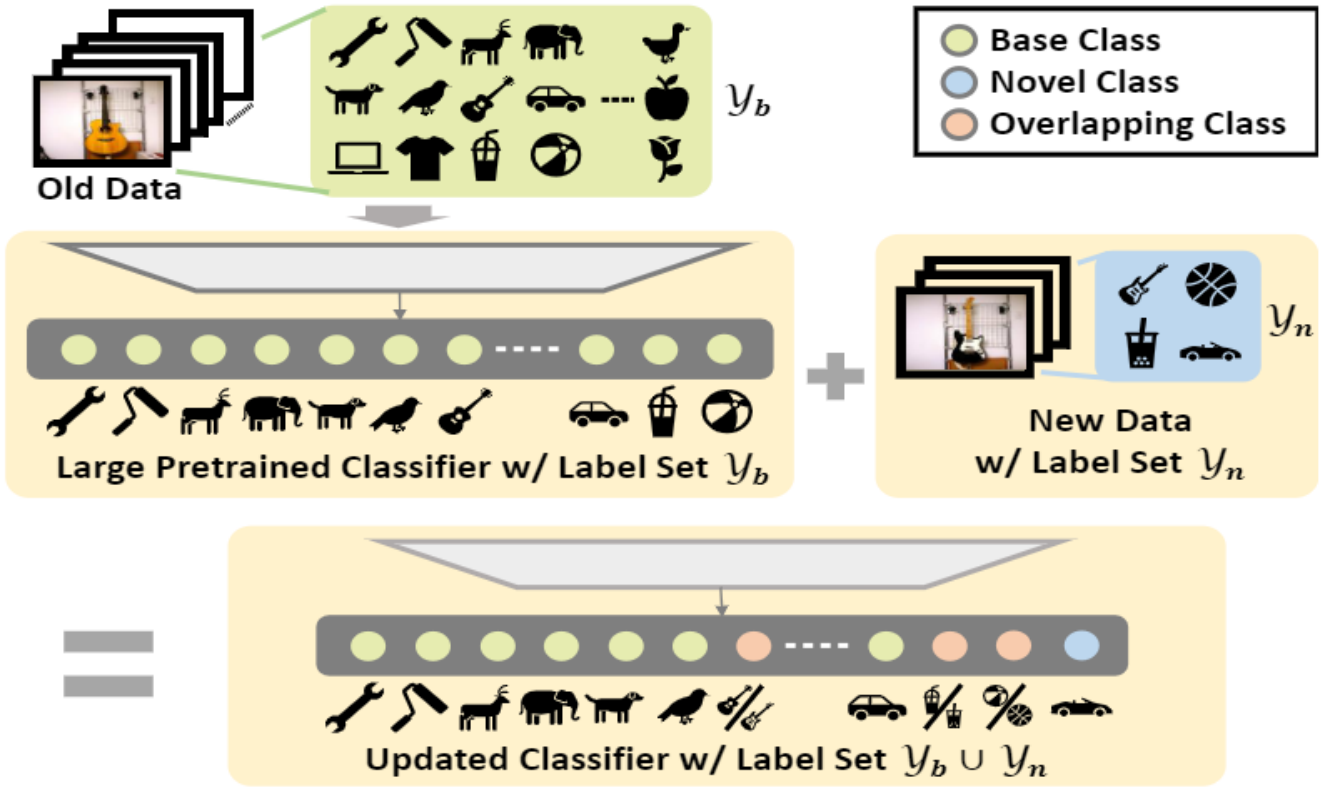
Table 2. The mIoU(%) of the last step on the ADE20K dataset for different overlapped continual learning scenarios. The **red** denotes the highest results and the **blue** denotes the second highest results.

Class-Incremental Learning with Strong Pre-trained Models

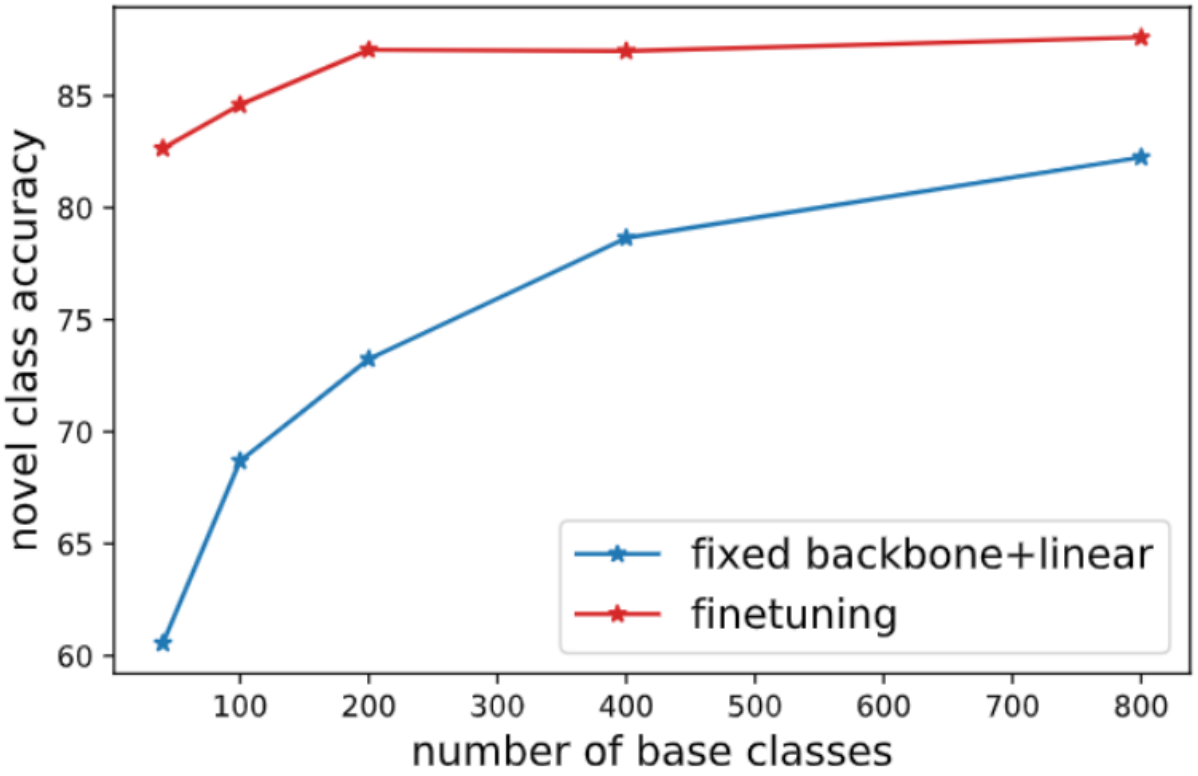
Tz-Ying Wu[1,2] Gurumurthy Swaminathan[1] Zhizhong Li[1]
Avinash Ravichandran[1] Nuno Vasconcelos[2] Rahul Bhotika[1] Stefano Soatto[1]
[1]AWS AI Labs [2]UC San Diego
{gurumurs,lzhizhon,ravinash,bhotikar,soattos}@amazon.com {tzw001,nuno}@ucsd.edu

Motivation

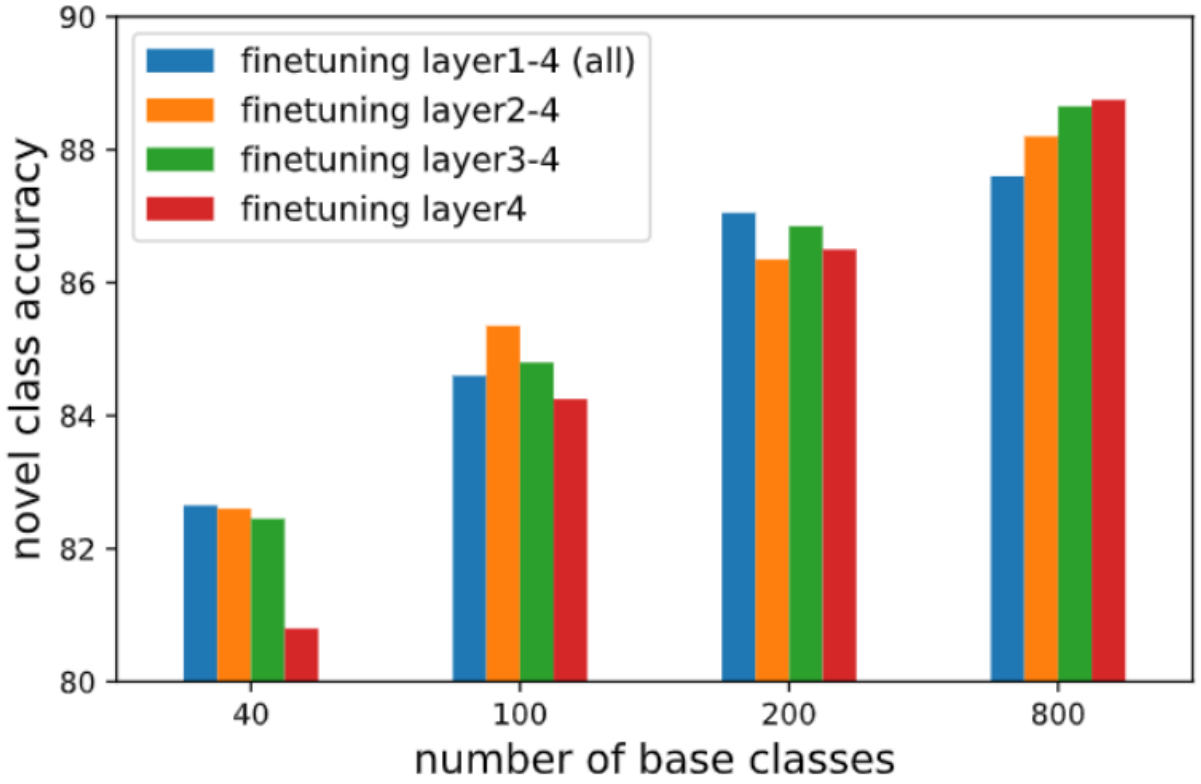
- Task Incremental Learning is not reality in some scenarios.
- Strong pre-trained models benefits Class-Incremental Learning



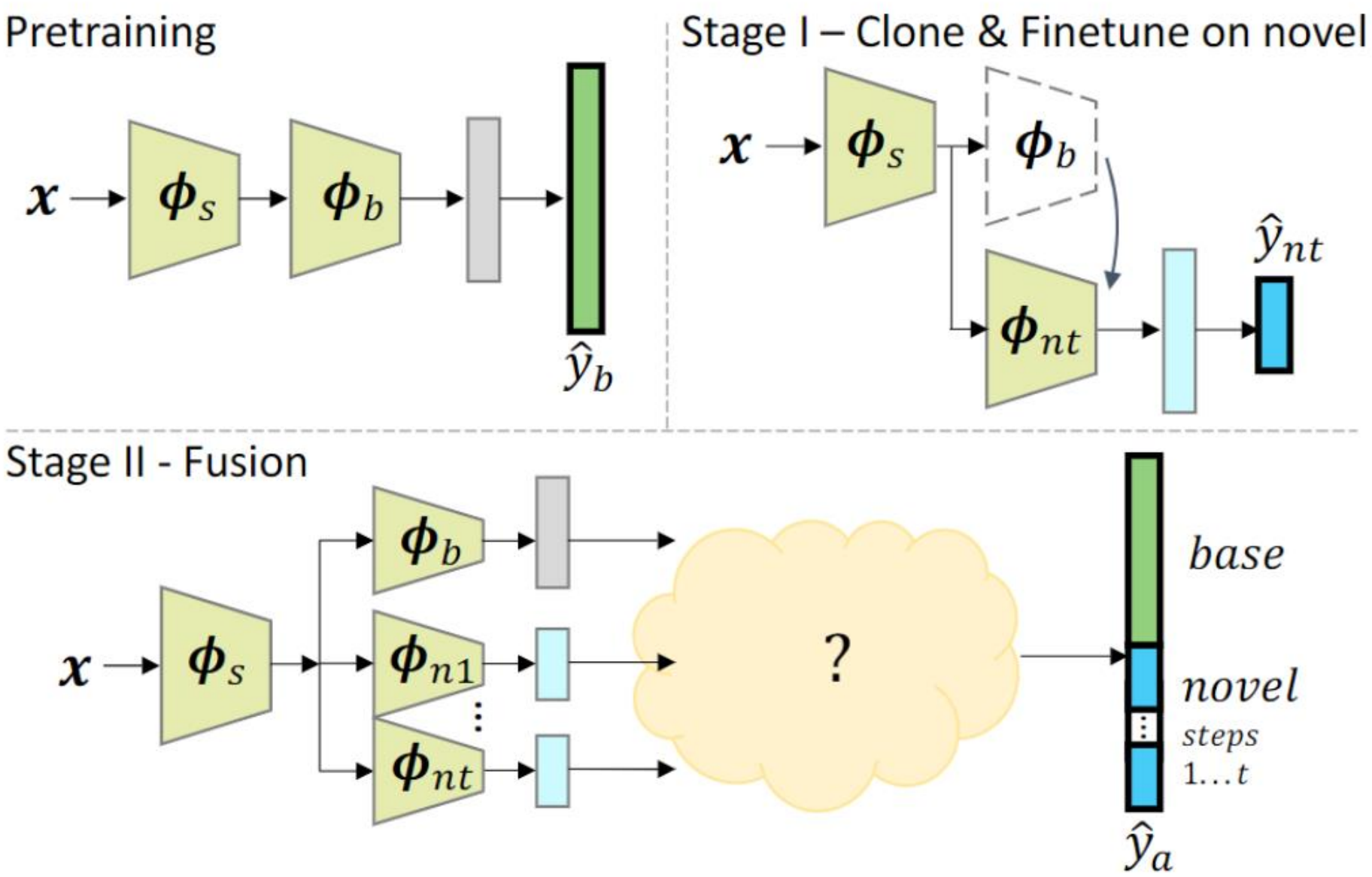
Observation

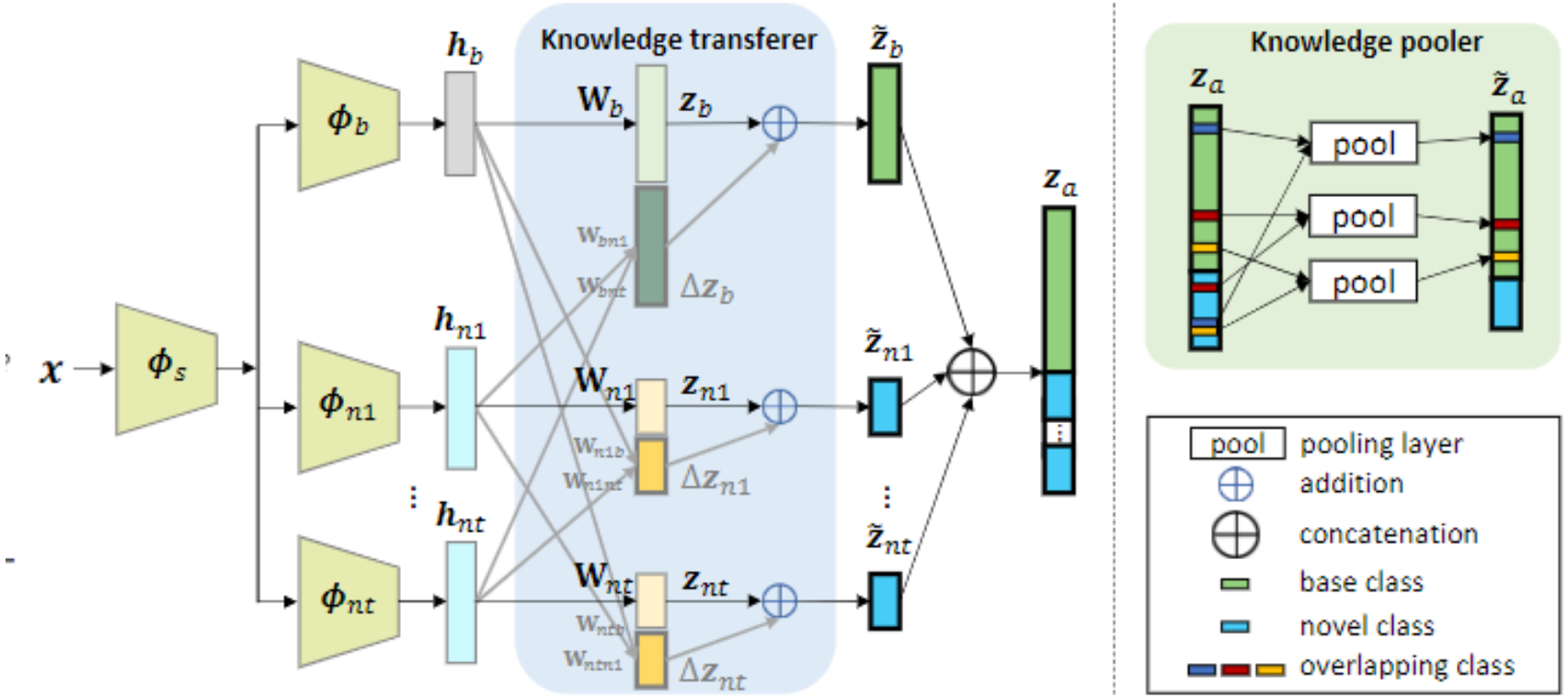


The comparison between fine-tuning and freezing representations.
Fixed with lots base classes performs better.

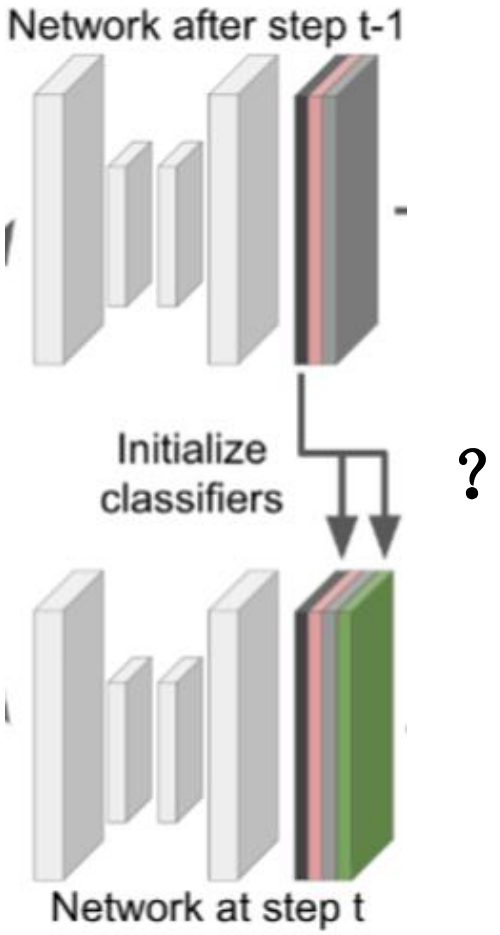


The novel class accuracy of fine-tuning each pre-trained model with different numbers of layers.


Pretraining
x
ϕ_s
ϕ_b
$\hat{y}_b$
Stage I – Clone & Finetune on novel
x
ϕ_s
ϕ_b
ϕ_{nt}
$\hat{y}_{nt}$
Stage II - Fusion
x
ϕ_s
ϕ_b
ϕ_{n1}
ϕ_{nt}
?
base
novel
steps
1…t
$\hat{y}_a$



(c) General score fusion network.



?

Method	ResNet10					ResNet18				
	# of params	Acc_{all}	Acc_{base}	Acc_{novel}	Acc_{avg}	# of params	Acc_{all}	Acc_{base}	Acc_{novel}	Acc_{avg}
fine-tuning	4.9M	4.18	0.01	87.63	43.82	11.2M	4.25	0.00	89.37	44.68
LwF [15]		9.50	5.54	88.53	47.04		9.50	5.46	90.30	47.88
iCaRL [23]		16.26	13.91	63.40	38.66		10.65	8.15	60.80	34.78
BiC [29]		30.30	27.55	85.20	56.38		31.50	28.75	86.60	57.68
WA [34]		51.33	52.33	31.40	41.87		54.79	55.17	47.20	51.19
DER w/o P [30]	9.8M	52.31	52.43	50.10	51.27	-	-	-	-	-
score fusion (ours) best-Acc_{all}	8.6M	**63.24**	63.77	52.67	58.22	19.6M	**69.45**	70.01	58.13	64.07
score fusion (ours) best-balanced		62.15	61.49	75.37	68.43		67.36	66.61	82.37	74.49
score fusion (ours) best-Acc_{avg}		58.90	57.73	82.40	**70.06**		65.83	64.85	82.50	**75.17**
score fusion (ours, fc-only) best-Acc_{all}	4.9M	62.65	63.56	44.53	54.05	11.2M	68.79	69.58	53.07	61.32
score fusion (ours, fc-only) best-balanced		61.01	60.81	65.07	62.94		66.76	66.50	71.83	69.17
score fusion (ours, fc-only) best-Acc_{avg}		57.91	57.24	71.57	64.40		65.89	65.49	73.77	69.63
joint learning (oracle)	4.9M	63.80	63.94	61.00	62.47	11.2M	70.32	70.43	68.20	69.32