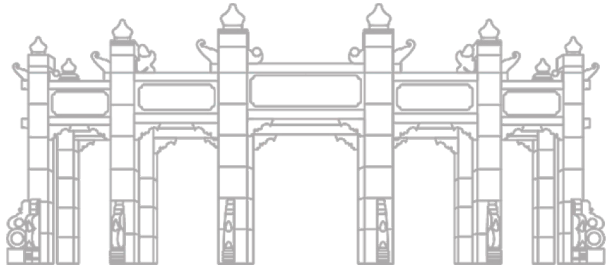
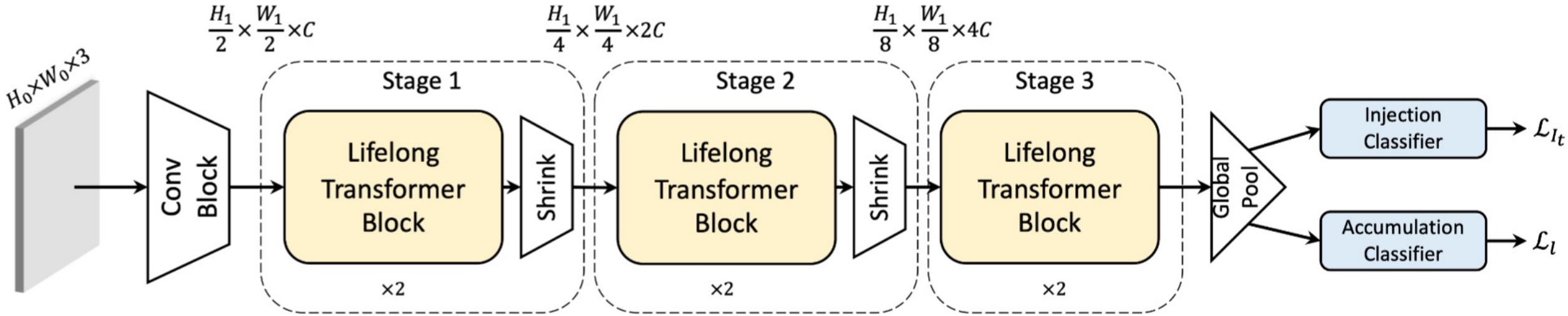
SYSU 2023
PRESENTATION SLIDE

中山大學
SUN YAT-SEN UNIVERSITY

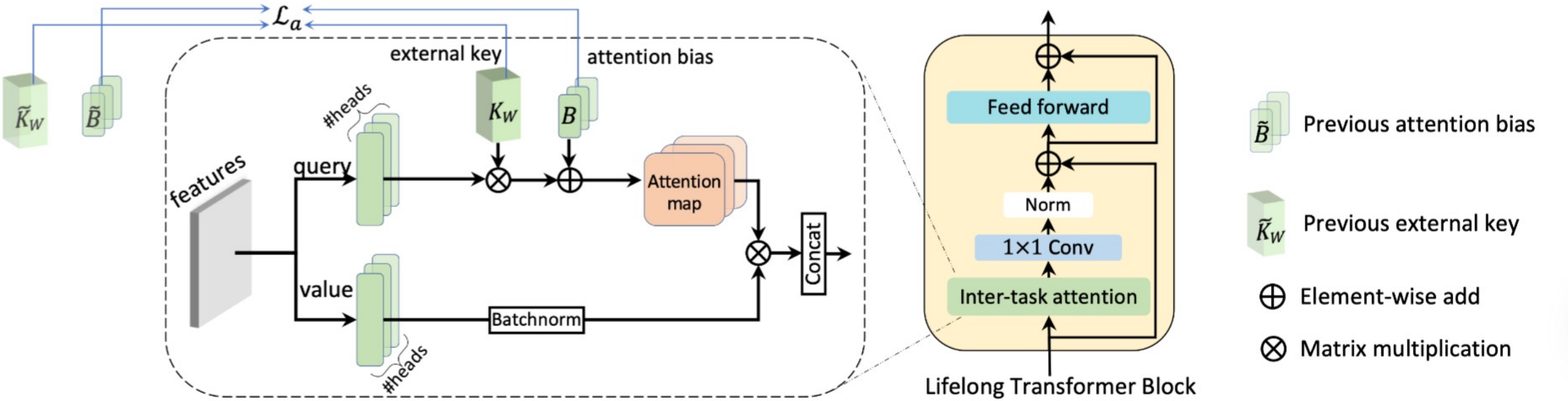
Continual Learning with Lifelong Vision Transformer

使用终身学习视觉transformer进行持续学习





(a) Overall architecture of LVT.

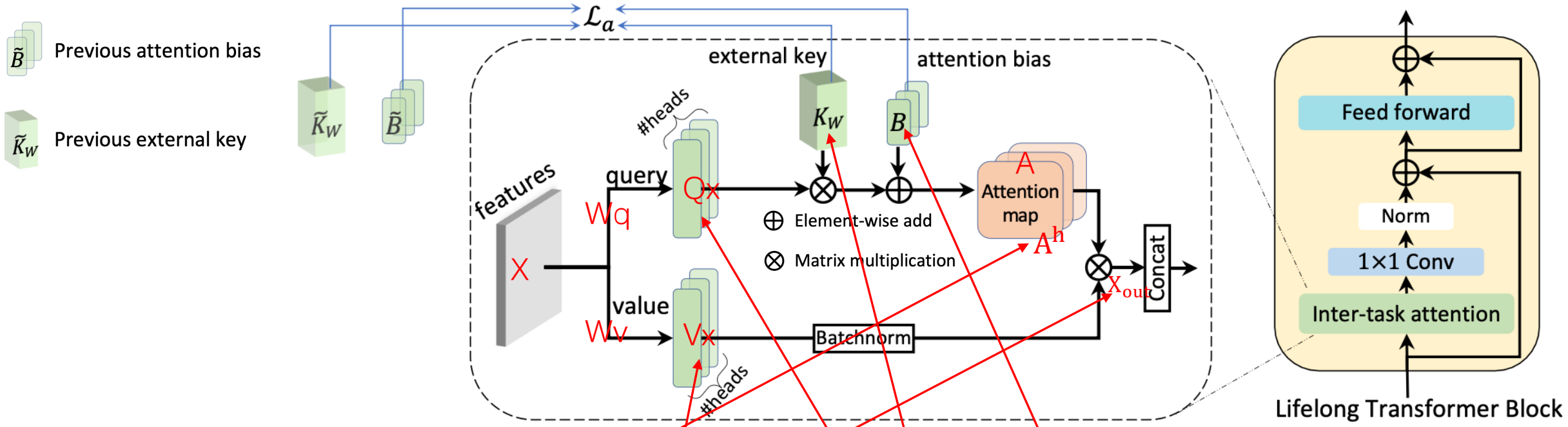


(b) Lifelong transformer block with inter-task attention mechanism.

$$\mathcal{L}_a = \left\| \nabla_{\widetilde{K}_W} \mathcal{L}_{I_t} \odot \left(K_W - \widetilde{K}_W \right) \right\|_1 + \left\| \nabla_{\widetilde{B}} \mathcal{L}_{I_t} \odot \left(B - \widetilde{B} \right) \right\|_1, \quad (2)$$

⊙：Hadamard卷积
||·||：L1正则化
L_It：式（3）的交叉熵损失

∇Kw_L_It、∇B_L_It：通过最后一项任务的平均损失梯度计算出的重要性，分别与Kw、B相关



$$A^h = \text{Softmax}\left(\frac{\text{Norm}(Q_X^h (K_W^h)^\top) + B^h}{\sqrt{d/H}} \right),$$

$$X_{out}^h = A^h V_X^h, \quad h = 1, \ldots, H,$$

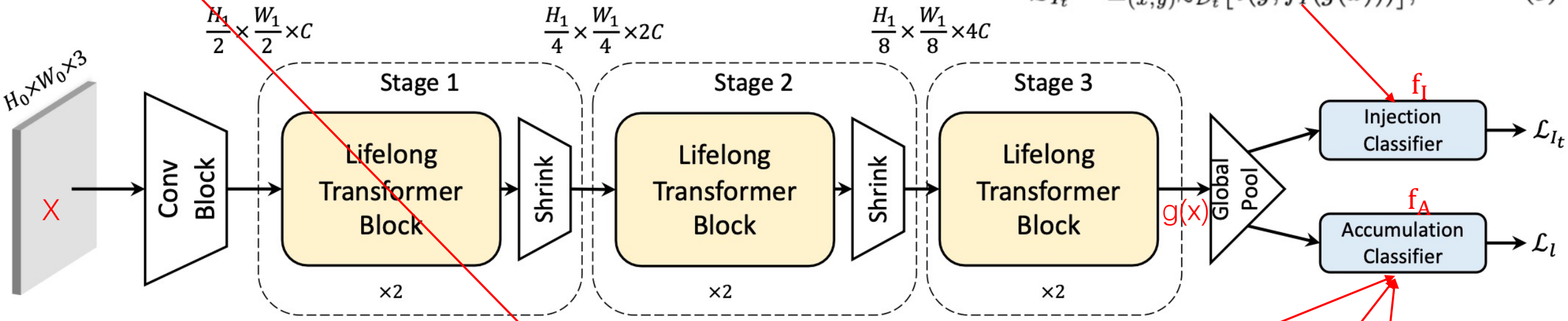
X：输入张量
Wq、Wv：线性transformer参数
B：加到注意力图上的偏置
H：多头注意力头的数量
d：K和Q的维度
Norm()：batch归一化函数

使用上标h来区别不同头的参数
A_h：某个注意力头算出来的注意力图

$$\mathcal{L} = \mathcal{L}_l + \mathcal{L}_{I_t} + \gamma \mathcal{L}_a, \qquad (7)$$

fI：添加分类器
l：交叉熵损失
g：LVT主干

$$\mathcal{L}_{I_t} = \mathbb{E}_{(x,y)\sim\mathcal{D}_t}\left[\ell(y, f_I(g(x)))\right], \qquad (3)$$



h_A：没有softmax操作的累加分类器
f_A：h_A之后再进行softmax操作

$$\mathcal{L}_r = \mathbb{E}_{(x',y')\sim\mathcal{M}}\left[\ell(y', f_A(g(x')))\right], \qquad (4)$$

$$z = h_A(g(x))$$

$$\mathcal{L}_d = \mathbb{E}_{(x',y',z')\sim\mathcal{M}}\left[D_{KL}\left(\mathrm{softmax}(z') \,||\, f_A(g(x'))\right)\right], \qquad (5)$$

$$\mathcal{L}_{A_t} = \mathbb{E}_{(x,y)\sim\mathcal{D}_t}\left[\ell(y, f_A(g(x)))\right]$$

α、β：用于平衡知识合并的系数
t：目前观察到的任务的数量
r(t)：一个和t有关的单调减函数

$$\mathcal{L}_l = \alpha\,\mathcal{L}_r + \beta\,\mathcal{L}_d + r(t)\mathcal{L}_{A_t}, \qquad (6)$$

$$\rho(x) = \frac{e^{z^c}}{\sum_{i=1}^{|C|} e^{z^i}},$$

$$x \in \{\hat{x} | (\hat{x}, \hat{y}) \in \mathcal{D}_t, \hat{y} = y_c\}$$

$$z = h_A(g(x))$$

$$c \in \mathcal{C}_t$$

C：迄今已观察到的类集合
Ct：任务t的类集合
c：任务t的某个类

z^i：z的第i个元素
y_c：类c对应的标签
z^c：应该是z中识别为类c的元素

Memory Buffer	Method	#Paras	5 splits		10 splits		20 splits	
			Class-IL	*Task-IL*	*Class-IL*	*Task-IL*	*Class-IL*	*Task-IL*
–	Joint	11.2	70.21±0.15	85.25±0.29	70.21±0.15	91.24±0.27	71.25±0.22	94.02±0.33
	SGD	11.2	17.27±0.14	42.24±0.33	8.62±0.09	34.40±0.53	4.73±0.06	40.83±0.46
200	ER [62]	11.2	21.94±0.83	62.41±0.93	14.23±0.12	67.57±0.68	9.90±1.67	70.82±0.74
	GEM [47]	11.2	19.73±0.34	57.13±0.94	13.20±0.21	62.96±0.67	8.29±0.18	66.28±1.49
	AGEM [15]	11.2	17.97±0.26	53.55±1.13	9.44±0.29	55.04±0.87	4.88±0.09	41.30±0.56
	iCaRL [61]	11.2	30.12±2.45	55.70±1.87	22.38±2.79	60.81±2.48	12.62±1.43	62.17±1.93
	FDR [9]	11.2	22.84±1.49	63.75±0.49	14.85±2.76	65.88±0.60	6.70±0.79	59.13±0.73
	GSS [4]	11.2	19.44±2.83	56.11±1.50	11.84±1.46	56.24±0.98	6.42±1.24	51.64±2.89
	DER++ [10]	11.2	27.46±1.16	62.55±2.31	21.76±0.78	59.54±0.77	15.16±1.53	61.98±0.91
	HAL [14]	22.4	13.21±1.24	35.61±2.95	9.67±1.67	37.49±2.16	5.67±0.91	53.06±2.87
	ERT [11]	11.2	21.61±0.87	54.75±1.32	12.91±1.46	58.49±3.12	10.14±1.96	62.90±2.72
	RM [7]	11.2	32.23±1.09	62.05±0.62	22.71±0.93	66.28±0.60	15.15±2.14	68.21±0.43
	LVT (ours)	**8.9**	**39.68**±1.36	**66.92**±0.40	**35.41**±1.28	**72.80**±0.49	**20.63**±1.14	**73.41**±0.67
500	ER [62]	11.2	27.97±0.33	68.21±0.29	21.54±0.29	74.97±0.41	15.36±1.15	74.97±1.44
	GEM [47]	11.2	25.44±0.72	67.49±0.91	18.48±1.34	72.68±0.46	12.58±2.15	**78.24**±0.61
	AGEM [15]	11.2	18.75±0.51	58.70±1.49	9.72±0.22	58.23±0.64	5.97±1.13	59.12±1.57
	iCaRL [61]	11.2	35.95±2.16	64.40±1.59	30.25±1.86	71.02±2.54	20.05±1.33	72.26±1.47
	FDR [9]	11.2	29.99±2.23	69.11±0.59	22.81±2.81	74.22±0.72	13.10±3.34	73.22±0.83
	GSS [4]	11.2	22.08±3.51	61.77±1.52	13.72±2.64	56.32±1.84	7.49±4.78	57.42±1.61
	DER++ [10]	11.2	38.39±1.57	70.74±0.56	36.15±1.10	73.31±0.78	21.65±1.44	70.55±0.87
	HAL [14]	22.4	16.74±3.51	39.70±2.53	11.12±3.80	41.75±2.17	9.71±2.91	55.60±1.83
	ERT [11]	11.2	28.82±1.83	62.85±0.28	23.00±0.58	68.26±0.83	18.42±1.92	73.50±0.82
	RM [7]	11.2	39.47±1.26	69.27±0.41	32.52±1.53	73.51±0.89	23.09±1.72	75.06±0.75
	LVT (ours)	**8.9**	**44.73**±1.19	**71.54**±0.93	**43.51**±1.06	**76.78**±0.71	**26.75**±1.29	78.15±0.42

Table 1. Results (overall accuracy %) on CIFAR100 benchmark which is averaged over five runs. #Paras means the number of parameters in the model, which is counted by million.

Memory Buffer	Method	#Paras	TinyImageNet		#Paras	ImageNet100	
			Class-IL	*Task-IL*		*Class-IL*	*Task-IL*
–	Joint	11.2	59.36±0.19	81.95±0.15	11.2	73.82±0.23	81.58±0.31
	SGD	11.2	7.87±0.24	18.31±0.63	11.2	8.72±0.37	21.32±0.61
200	ER [62]	11.2	8.79±0.21	39.16 ±2.14	11.2	9.58±0.34	36.24±1.69
	AGEM [15]	11.2	8.28±0.15	23.79±0.11	11.2	9.27±0.08	25.20±0.35
	iCaRL [61]	11.2	8.64±0.78	28.41±1.53	11.2	12.59±0.68	33.75±1.81
	FDR [9]	11.2	8.77±0.82	40.15±0.67	11.2	10.08±0.36	37.80±0.91
	DER++ [10]	11.2	11.16±0.95	40.97±1.16	11.2	11.92±0.12	31.96±1.65
	ERT [11]	11.2	10.85±0.24	39.54±1.90	11.2	13.51±1.13	36.94±1.54
	RM [7]	11.2	13.58±1.07	41.96±1.28	11.2	16.76±0.84	35.18±1.43
	LVT (ours)	**9.0**	**17.34**±1.13**(+3.76)**	**46.15**±1.21**(+4.19)**	**9.4**	**19.46**±1.06**(+2.70)**	**41.78**±2.03**(+3.98)**
500	ER [62]	11.2	10.15±0.32	50.11±0.53	11.2	11.68±0.25	42.04±0.47
	AGEM [15]	11.2	9.67±0.18	26.79±0.81	11.2	10.92±0.16	34.22±0.68
	iCaRL [61]	11.2	10.69±1.53	35.89±2.47	11.2	16.44±1.35	36.89±0.72
	FDR [9]	11.2	10.58±0.22	49.91±0.78	11.2	11.78±0.40	42.60±0.64
	DER++ [10]	11.2	19.33±1.41	51.90±0.62	11.2	14.52±1.86	35.46±0.66
	ERT [11]	11.2	12.13±0.36	50.87±0.49	11.2	20.42±1.13	41.56±1.78
	RM [7]	11.2	18.96±1.34	52.08±0.84	11.2	14.56±2.64	38.66±2.47
	LVT (ours)	**9.0**	**23.97**±1.27**(+4.64)**	**57.39**±0.75**(+5.31)**	**9.4**	**26.32**±1.67**(+5.90)**	**47.84**±1.33**(+5.24)**

Table 2. Results (overall accuracy %) on TinyImageNet and ImagNet100, which are averaged over three runs. #Paras means the number of parameters in the model, which is counted by million. The **green numbers** represent gains.

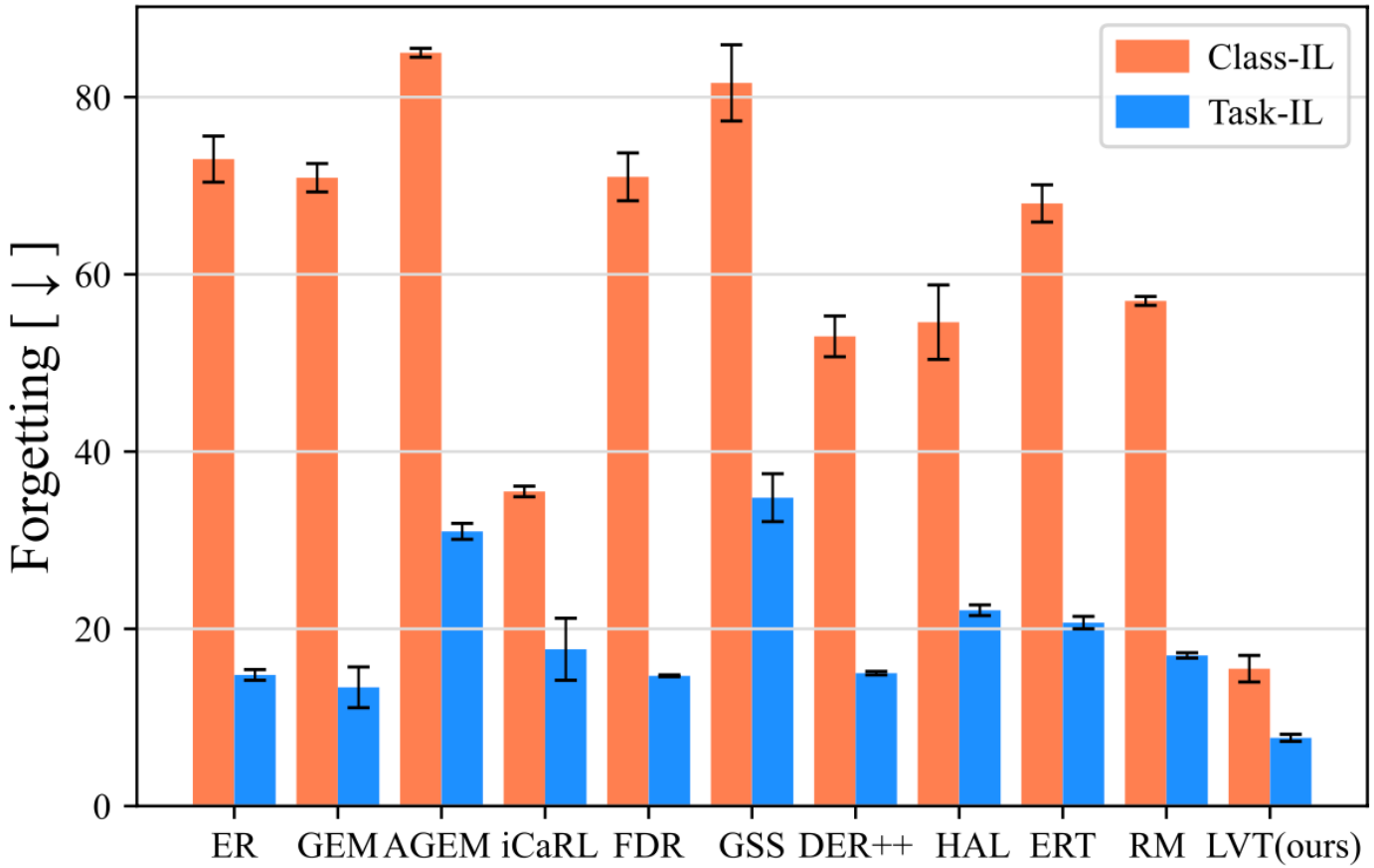


Figure 3. Forgetting results (%) on CIFAR100 (lower is better).

Method	**#Paras**	**Accuracy [↑]**		**Forgetting [↓]**	
		Class-IL	*Task-IL*	*Class-IL*	*Task-IL*
ViT [21]	16.2	13.19	54.53	70.16	24.79
LeViT [25]	10.9	31.84	72.76	52.93	14.67
CoaT [85]	10.3	25.44	66.15	58.01	17.28
CCT [32]	**3.9**	24.50	71.37	66.17	20.20
ResNet18 [33]	11.2	36.98	73.23	47.43	15.36
LVT (ours)	8.9	**43.51**	**76.78**	**15.54**	**7.76**

Table 3. Comparison with vision transformer and CNN architectures for continual learning.

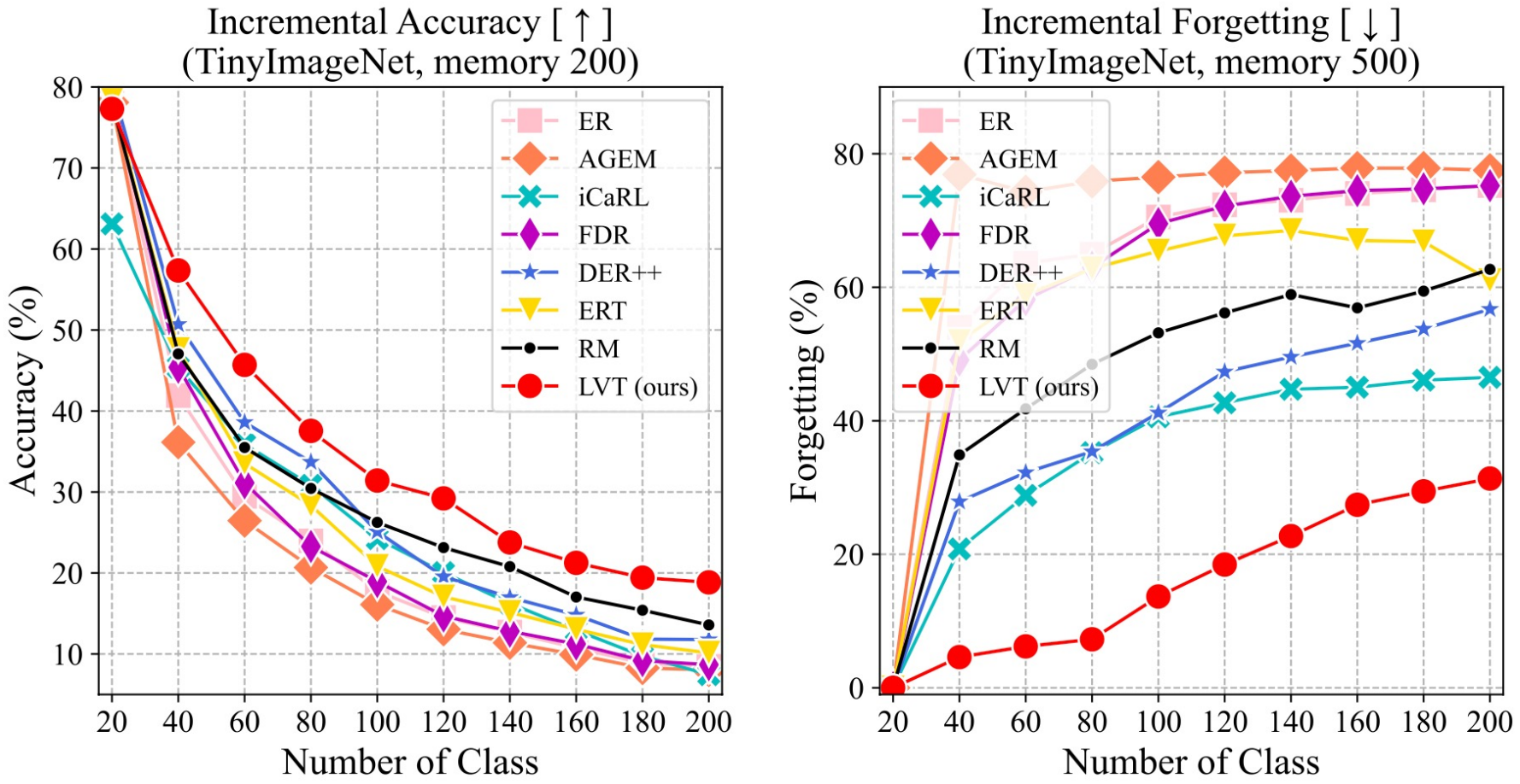
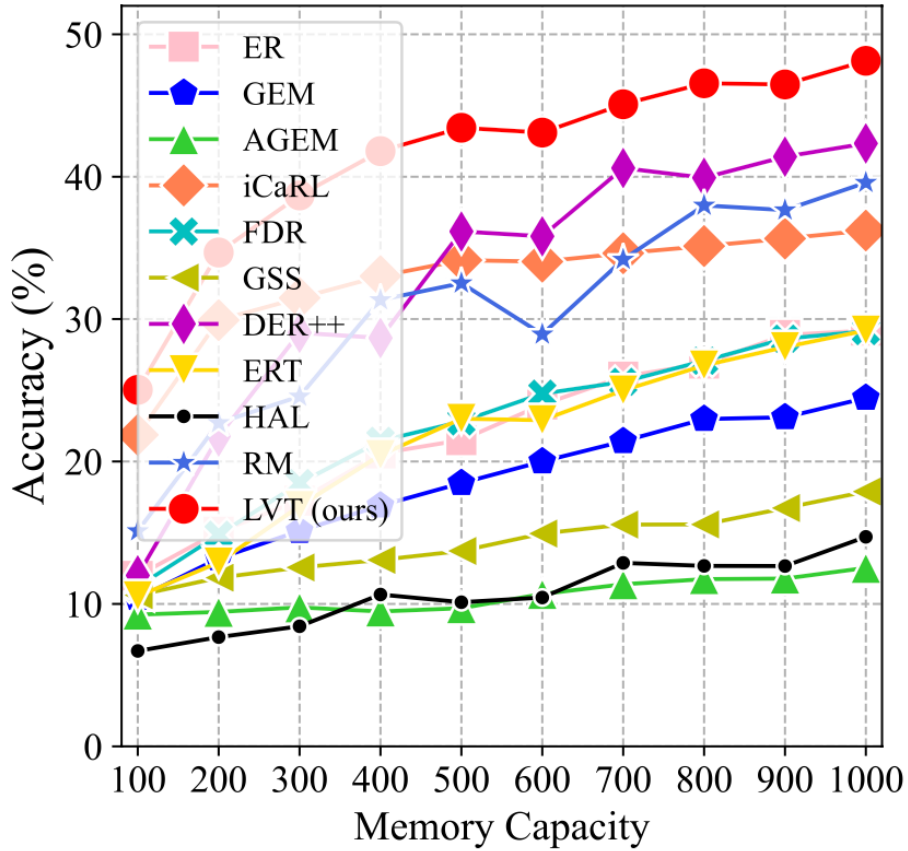


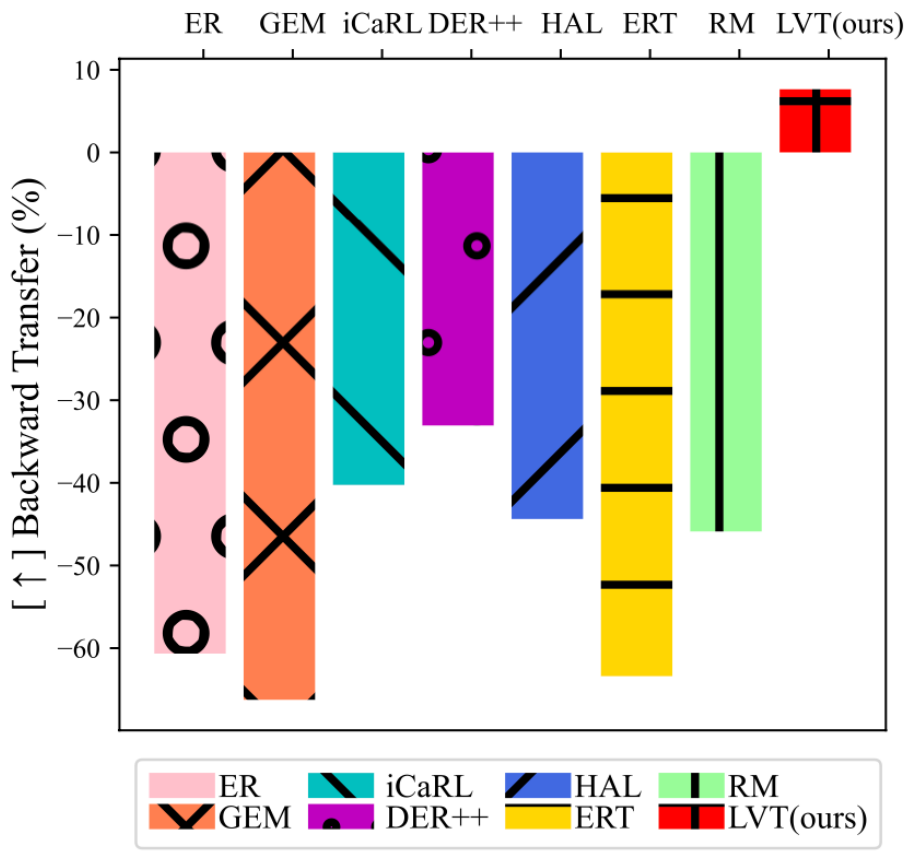
Figure 4. Incremental performance evaluated on all tasks observed so far. [↑] higher is better, [↓] lower is better.

Module				CIFAR100		TinyImageNet	
IT-*att*	f_I	f_A	ρ	*Class-IL*	*Task-IL*	*Class-IL*	*Task-IL*
	√	√	√	36.93	73.52	19.35	52.14
√		√	√	39.76	74.78	20.03	54.34
√	√		√	38.42	75.49	18.85	55.07
√	√	√		40.25	73.71	21.16	54.42
√	√	√	√	**43.51**	**76.78**	**23.97**	**57.39**

Table 4. Ablation study on each component of LVT. IT-att represents the transformer with inter-task attention mechanism; f_I and f_A denotes the injection classifier and accumulation classifier respectively; And ρ denotes the confidence-ware memory update.



(a) Sensitivity analysis regarding the memory capacity.



(b) Backward transfer (BWT) analysis under Class-IL setting.

Figure 5. Analyses for memory capacity and backward transfer.

$$\text{BWT}=\frac{1}{T-1}\sum_{t=1}^{T-1}\left(a_{T,t}-a_{t,t}\right)$$

中山大學
SUN YAT-SEN UNIVERSITY

谢谢观看