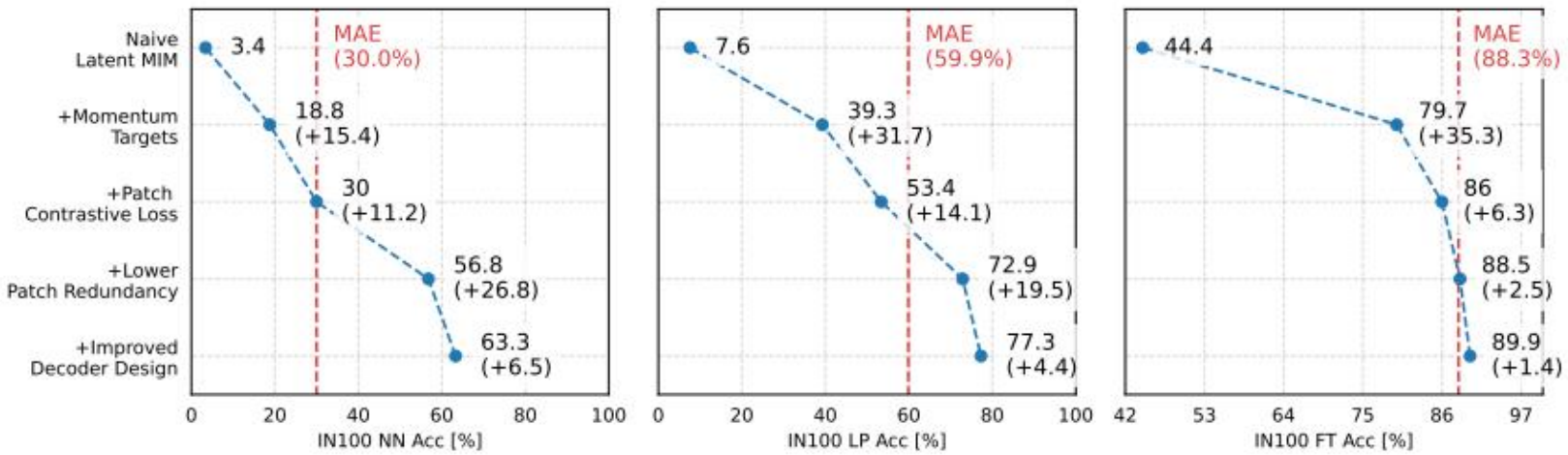
Towards Latent Masked Image Modeling for Self-Supervised Visual Representation Learning

面向用于自监督视觉表征学习的潜在掩码图像建模

背景



潜在掩码图像建模所面临的重大挑战

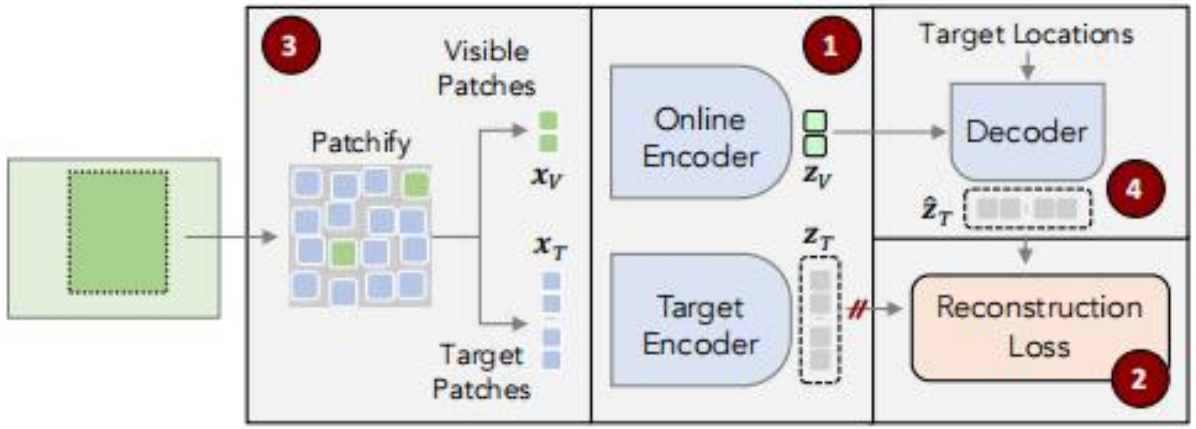
1、联合优化可见区域和被掩码区域表征会导致表征崩塌

2、损失函数的选择和设计

3、控制可见区域和目标区域之间的冗余性

4、解码器设计必须精心设计

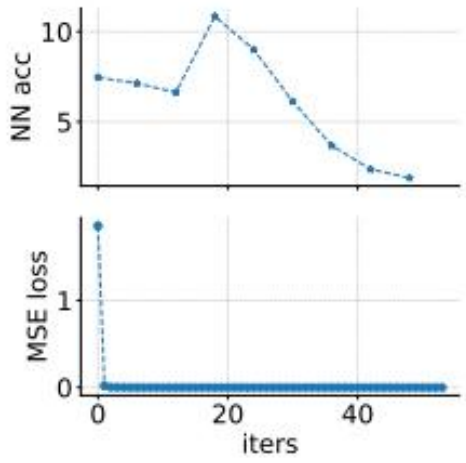
方法架构



Challenges of Latent MIM

1. Joint online/target optimization results in representation collapse.
2. Direct reconstruction is conducive to trivial solutions.
3. Effective learning requires low correlation between visible and target patches.
4. Decoder must be designed for latent reconstruction with latent conditioning.

表征崩溃

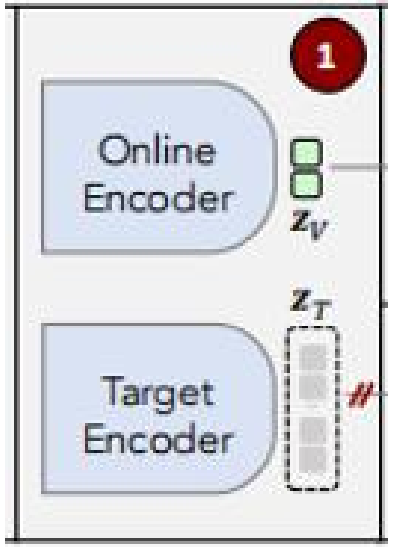


解决方案：

1、独立目标编码器

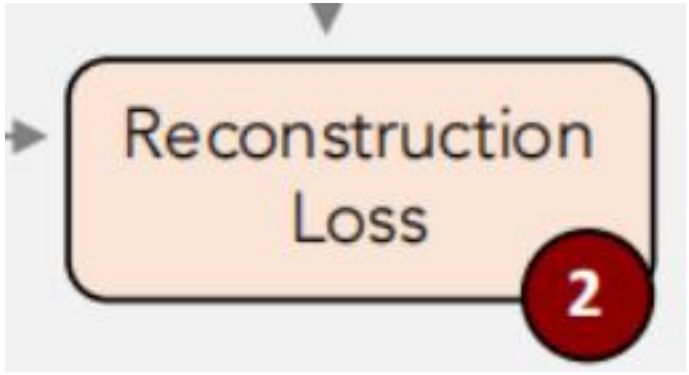
2、带停止梯度的权重共享

3、动量编码器



Method	Sim	NN	LP	Ft
Naive Latent-MIM	1.00	0.9	7.9	nan
No weight sharing	0.96	3.4	7.6	44.4
Weight sharing & stop-grad	0.99	11.3	26.6	56.9
Momentum targets	**0.50**	18.8	39.3	79.7
MAE [13]	0.67	**30.0**	**70.4**	**88.3**

损失函数



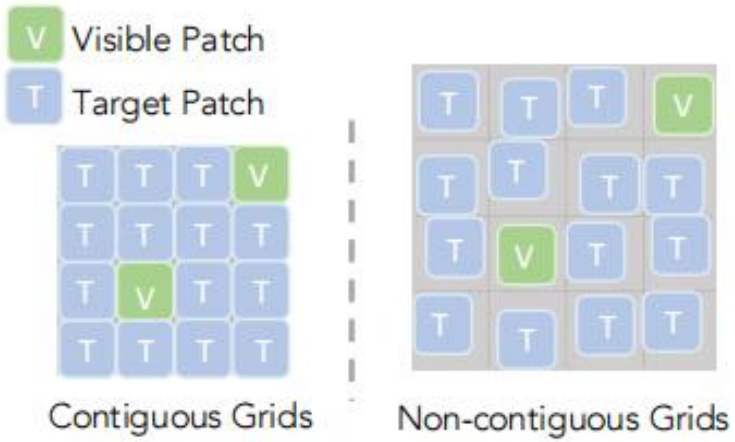
$$\Delta^k_{L2} = \|\hat{\boldsymbol{z}}_k - \boldsymbol{z}_k\|_2^2, \quad \Delta^k_{L1} = \|\hat{\boldsymbol{z}}_k - \boldsymbol{z}_k\|_1,$$

$$\Delta^k_{Huber} = \begin{cases} \frac{1}{2}\Delta^k_{L2} & \text{if } \Delta_{L2} < \delta^2 \\ \delta \cdot \left(\Delta^k_{L1} - \delta/2\right) & \text{otherwise.} \end{cases}$$

$$\Delta^k_{PatchDisc} = -\tau \log \frac{\exp\left(-\frac{1}{\tau}\mathtt{sim}\left(\hat{\boldsymbol{z}}_k, \boldsymbol{z}_k\right)\right)}{\sum_{l \in \mathcal{T}} \exp\left(-\frac{1}{\tau}\mathtt{sim}\left(\hat{\boldsymbol{z}}_k, \boldsymbol{z}_l\right)\right)}, \quad \mathtt{sim}(\hat{\boldsymbol{z}}, \boldsymbol{z}) = \frac{\hat{\boldsymbol{z}}^T \boldsymbol{z}}{\|\hat{\boldsymbol{z}}\|\|\boldsymbol{z}\|}$$

Loss	NN	LP	Ft
MSE	18.8	39.3	79.7
L1	15.2	36.7	81.7
Huber	24.3	46.4	82.1
PatchDisc	**30.0**	53.4	86.0
MAE [13]	**30.0**	**59.9**	**88.3**

相邻图像块之间的语义相关性



解决方案

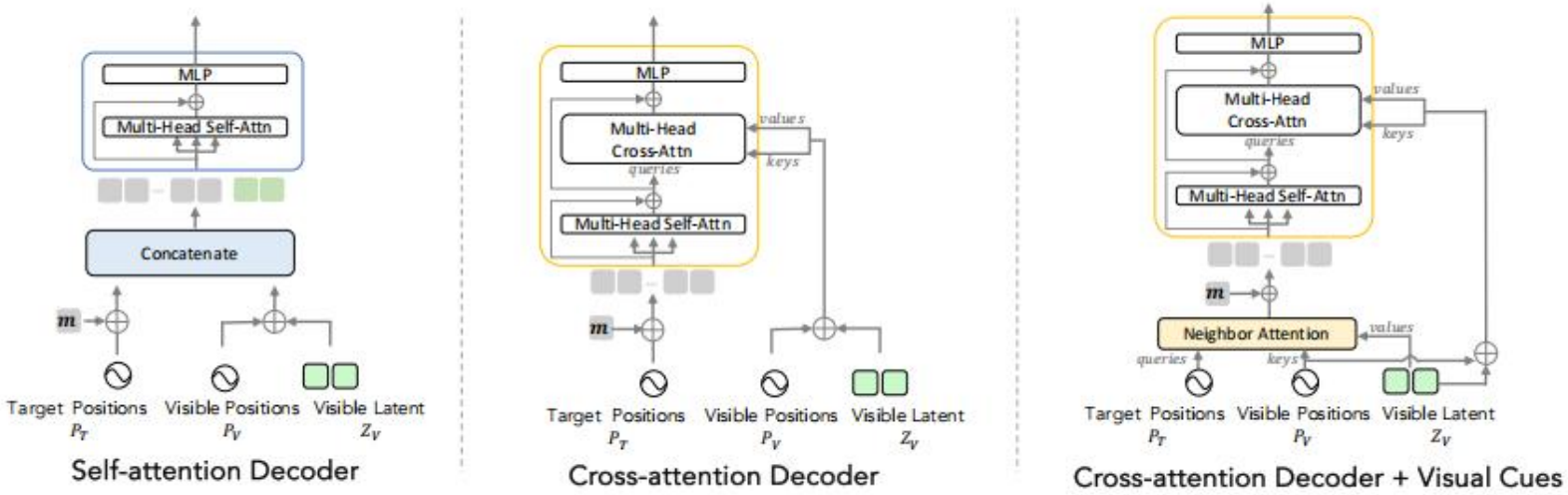
1、非连续网格

2、图像块相似性约束

Mask	Gap	Sim.	NN	LP	Ft
0.75	0	✗	30.0	53.4	86.0
0.9	0	✗	53.3	70.4	87.6
0.9	4	✗	54.6	71.2	88.0
0.9	4	✓	**56.8**	**72.9**	**88.5**
MAE [13]			30.0	59.9	88.3

$$\mathcal{R} = \left(\gamma - \mathbb{E}_{i,j\in\mathcal{T}}\left[\mathtt{sim}\left(\hat{\boldsymbol{z}}_i, \hat{\boldsymbol{z}}_j\right)\right]\right)^2 + \left(\gamma - \mathbb{E}_{i,j\in\mathcal{V}}\left[\mathtt{sim}\left(\boldsymbol{z}_i, \boldsymbol{z}_j\right)\right]\right)^2,$$

解码器的设计



$$\boldsymbol{m}_i = \boldsymbol{m} + \boldsymbol{p}_t + Softmax_t\left(P_{\mathcal{T}}P_{\mathcal{V}})^T\right) Z_{\mathcal{V}}$$

Decoder	Proj.	Depth	NN	LP	Ft
Self-attn	none	3	56.8	72.9	88.5
Self-attn	mlp	3	59.7	76.8	89.4
Cross-attn	mlp	3	61.1	77.0	89.4
Cross-attn w/ Vis Cues	mlp	3	**63.3**	**77.3**	**89.9**
Cross-attn w/ Vis Cues	mlp	8	35.9	65.8	86.2
MAE [13]		8	30.0	59.9	88.3

实验

Method	Epochs	NN	LP
Low-level			
MAE [13]	1600	12.2	67.8
SimMIM [28]	800	-	56.7
Latent			
ConMIM [29]	800	-	39.3
data2vec [2]	800	25.7	60.3
Latent MIM	800	**50.1**	**72.0**

Method	epochs	$\mathcal{J}\&\mathcal{F}_m$	$\mathcal{J}_m$	$\mathcal{F}_m$
MAE	1600	57.5	54.8	60.2
data2vec	800	28.5	27.8	29.2
Latent MIM	800	**65.5**	**63.1**	**68.0**

Method	Caltech101 [11]	DTD [8]	Oxford Flowers [17]	Oxford Pets [19]	Stanford Cars [15]	SUN397 [26]	UCF101 [20]
MAE	80.5± 0.4	51.8± 2.2	62.7± 1.8	60.2± 3.6	7.8± 1.0	21.9± 0.4	53.2± 0.4
data2vec	76.6± 2.1	52.0± 2.1	74.1± 2.2	58.5± 2.0	8.8± 0.4	29.0± 0.5	52.6± 1.2
Latent MIM	**89.2**± 1.0	**55.9**± 2.0	**84**± 1.0	**79.8**± 2.6	**16.3**± 2.0	**48.4**± 0.6	**75.8**± 2.8

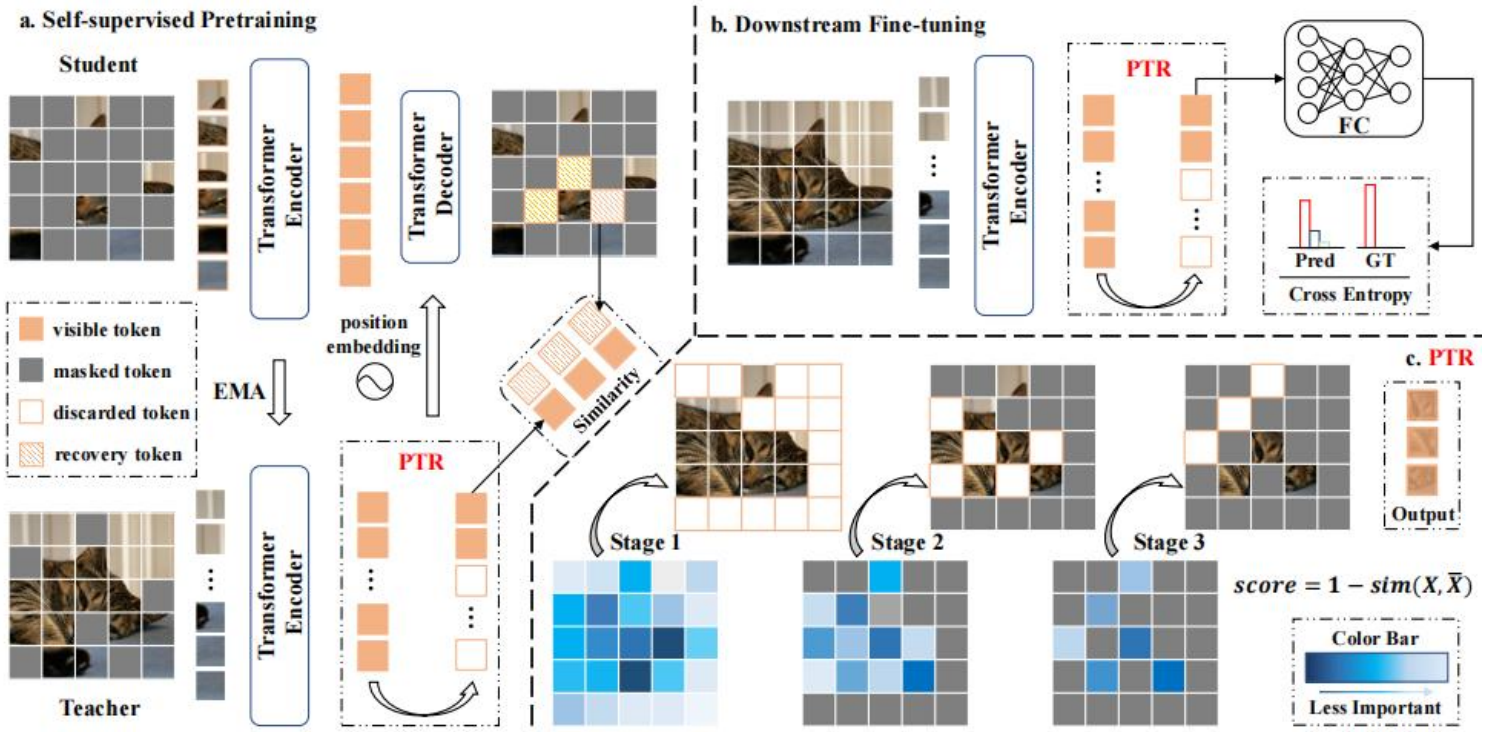
Progressively Compressed Autoencoder for Self-Supervised Representation Learning

面向用于自监督视觉表征学习的潜在掩码图像建模

背景

掩码图像建模从可见块中恢复所有被掩码的块来驱动学习，但是同一图像的块之间高度相关，重建所有的掩码块是冗余的。

方法架构

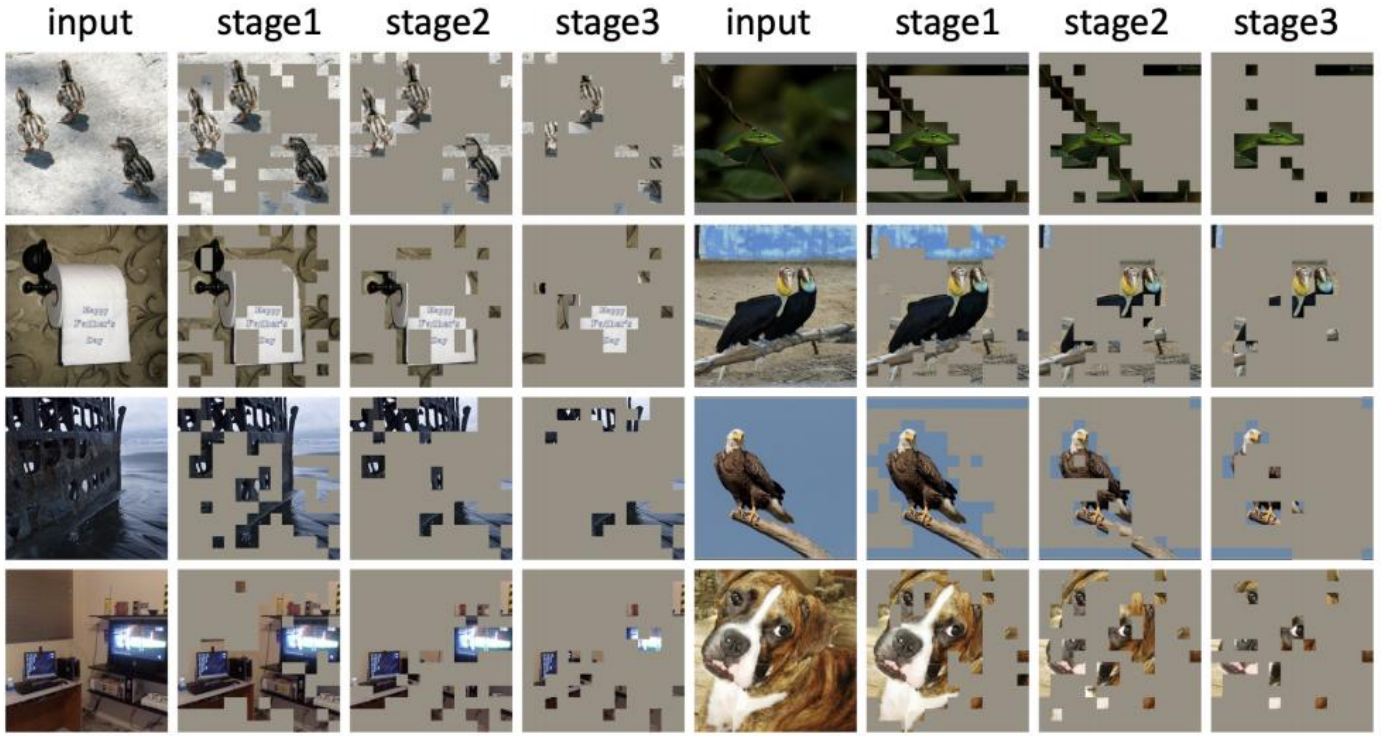


$$\max_{\theta,\phi,\omega} \mathbb{E}_{x\sim\mathcal{X}} \mathcal{M}\{\bar{f}[(1-M)\odot x;k], g_\phi\{h[f_\theta(M\odot x),\omega+p]\}\}$$

$$S(x) = rank(\mathcal{M}(x, \sum_j^{N_i} x^j/N_i)) \in \mathbb{R}^{N_i},$$

$$T(x;k) = gather[x, S(x) < \lfloor N_i * k \rfloor] \in \mathbb{R}^{\lfloor N_i * k \rfloor \times D}$$

效果图



实验

Methods	Epochs	Object Detection			Instance Segmentation		
		AP^b	AP^b_{50}	AP^b_{75}	AP^m	AP^m_{50}	AP^m_{75}
Supervised methods:							
DeiT	300	46.9	68.9	51.0	41.5	65.5	44.4
Contrastive learning methods:							
MoCo v3	300	45.5	67.1	49.4	40.5	63.7	43.4
DINO	400	**46.8**	68.6	50.9	**41.5**	65.3	44.5
Masked Image Modeling methods:							
BEiT	300	39.5	60.6	43.0	35.9	57.7	38.5
BEiT	800	42.1	63.3	46.0	37.8	60.1	40.6
MAE	300	45.4	66.4	49.6	40.6	63.4	43.7
MAE	1600	48.4	69.4	53.1	42.6	66.1	45.9
PCAE	300	47.6	68.3	52.4	42.0	65.3	45.6
PCAE	800	**48.8**	69.6	53.6	**43.1**	66.9	46.4

Method	Epochs	Forwards	GPU Days	Accuracy
Train from Scratch	300	1	-	81.8
MoCo v3	300	2	-	83.2
DINO	400	12	122.5	83.3
BEiT†	300	1	-	83.0
CAE†	300	2	-	83.6
SimMIM	800	1	54.1	**83.8**
MAE	300	1	15.4	82.9
MAE	1600	1	82.1	83.6
PCAE	300	2	10.1	83.6
PCAE	800	2	26.9	**83.9**

实验

Exp id	RD	drop case	GPUd	Acc.
Exp id = 1	✓	1	3.3	81.0
Exp id = 2	×	1	3.3	**82.3**
Exp id = 3	×	2	3.7	82.0
Exp id = 4	×	3	3.4	82.0
Exp id = 5	×	4	3.1	81.5
Exp id = 6	×	5	4.8	80.7