Supplementary Materials for On Data Scaling in Masked Image Modeling

Zhenda Xie[13], Zheng Zhang[3†], Yue Cao[3†], Yutong Lin[23], Yixuan Wei[13], Qi Dai[3], Han Hu[3]
[1]Tsinghua University [2]Xi'an Jiaotong University [3]Microsoft Research Asia
{t-zhxie, zhez, yuecao, t-yutonglin, t-yixuanwei, qi.dai, hanhu}@microsoft.com

A. Hyper-parameters and training details

We illustrate the training details of pre-training and fine-tuning for different tasks and different models. Table 1 presents pre-training details. Table 2 presents the fine-tuning details on ImageNet-1K image classification. Table 3 presents the fine-tuning details on iNaturalist 2018. Table 4 presents the fine-tuning details on COCO dataset. Table 5 presents the fine-tuning details on ADE20K dataset.

Pre-training setting of all models	
Input size	192^2
Window size	12
Patch size	4
Mask patch size	32
Mask ratio	0.6
Training iterations	125,000 / 250,000 / 500,000
Batch size	2048
Optimizer	AdamW
Init. learning rate	4e-4
Weight decay	0.05
Adam ϵ	1e-8
Adam β	(0.9, 0.999)
Learning rate scheduler	Step
Step learning rate ratio	0.1
Step iterations	109,375 / 218,750 / 437,500
Warm-up iterations	6250
Gradient clipping	5.0
Stochastic depth	0.1
Rand crop scale	[0.67, 1]
Rand resize ratio	[3/4, 4/3]
Rand horizontal flip	0.5
Reconstruction target	Norm. with sliding window [1]
Norm. patch size	47

Table 1. Details and hyper-parameters for SimMIM pre-training.

Hyperparameters	SwinV2	
	S / B / L	H / g
Input size	224^2	
Window size	14	
Patch size	4	
Training epochs	100	50
Warm-up epochs	20	10
Layer decay	0.8 / 0.75 / 0.7	0.65
Batch size	2048	
Optimizer	AdamW	
Base learning rate	5e-3	
Weight decay	0.05	
Adam ϵ	1e-8	
Adam β	(0.9, 0.999)	
Learning rate scheduler	cosine	
Gradient clipping	5.0	
Stochastic depth	0.2	
Label smoothing	0.1	
Rand crop scale	[0.08, 1]	
Rand resize ratio	[3/4, 4/3]	
Rand horizontal flip	0.5	
Color jitter	0.4	
Rand augment	9 / 0.5	
Rand erasing prob.	0.25	
Mixup prob.	0.8	
Cutmix prob.	1.0	

Table 2. Details and hyper-parameters for ImageNet-1K fine-tuning.

B. Training dynamics of masked image modeling

We show the training curves and validation curves of different models trained by masked image modeling to better illustrate the training dynamics. In Figure 2, each row presents the training and validation loss curves for training with the same model but different dataset. The training loss is computed on its corresponding training dataset and the validation loss is computed on the ImageNet-1K validation set. We make the following observations: First, all models

The work is done when Zhenda Xie, Yutong Lin, and Yixuan Wei are interns at Microsoft Research Asia. † Project co-leaders.

Hyperparameters		SwinV2	
	Small(S)	Base(B)	Large(L)
Input size		224^2	
Window size		14	
Patch size		4	
Training epochs		100	
Warm-up epochs		20	
Layer decay	0.8	0.75	0.7
Batch size		2048	
Optimizer		AdamW	
Base learning rate		1.6e-2	
Weight decay		0.1	
Adam ϵ		1e-8	
Adam β		(0.9, 0.999)	
Learning rate scheduler		cosine	
Gradient clipping		5.0	
Stochastic depth		0.2	
Label smoothing		0.1	
Rand crop scale		[0.08, 1]	
Rand resize ratio		[3/4, 4/3]	
Rand horizontal flip		0.5	
Color jitter		0.4	
Rand augment		9 / 0.5	
Rand erasing prob.		0.25	
Mixup prob.		0.8	
Cutmix prob.		1.0	

Table 3. Details and hyper-parameters for iNaturalist 2018 fine-tuning.

Hyperparameters		SwinV2	
	Small(S)	Base(B)	Large(L)
Detector		Mask R-CNN	
Window size		14	
Patch size		4	
Training input size		(1024, 1024)	
Testing input size		(800, 1333)	
Training epochs		36	
Warm-up iterations		500	
Batch size		32	
Optimizer		AdamW	
Base learning rate		8e-5	
Weight decay		0.05	
Adam ϵ		1e-8	
Adam β		(0.9, 0.999)	
Learning rate scheduler		Step	
Step learning rate ratio		0.1	
Step epochs		(27, 33)	
Stochastic depth	0.1	0.1	0.2
Rand horizontal flip		0.5	
Scale Jittering		[0.1, 2.0]	

Table 4. Details and hyper-parameters for fine-tuning on the COCO dataset.

Hyperparameters		SwinV2	
	Small(S)	Base(B)	Large(L)
Architecture		UPerNet	
Window size		20	
Patch size		4	
Training input size		(640, 640)	
Test input size		(640, 2560)	
Slide test stride		(426, 426)	
Training iterations		80,000	
Warm-up iterations		750	
Layer decay	0.95	0.95	0.9
Batch size		32	
Optimizer		AdamW	
Base learning rate		[1e-4, 3e-4]	
Weight decay		0.05	
Adam ϵ		1e-8	
Adam β		(0.9, 0.999)	
Learning rate scheduler		Linear	
Stochastic depth		0.1	
Rand horizontal flip		0.5	
Scaling Jittering		[0.5, 2.0]	
Photo Metric Distortion		✓	

Table 5. Details and hyper-parameters for fine-tuning on the ADE20K dataset.

have the overfitting issues when using small datasets. Second, for the non-overfitting cases, the training and validation losses are similar using different sizes of datasets for training. In Figure 3, the training/validation loss curves of different models but using the same training dataset are presented at each row. We make the following observations: First, larger models have lower training losses than smaller models for all datasets. Second, the validation loss of the larger model is lower than the smaller model in the non-overfitting cases but higher than the smaller model in the over-fitting cases.

C. Visualization

To better understand the difference between overfitting and non-overfitting models, we visualize the reconstruction results of SwinV2-L that pre-trained on ImageNet1K(10%) and ImageNet1K(100%). Figure. 4 shows the reconstruction results on the training images from ImageNet1K(10%) dataset that are jointly contained by the two models, and Figure. 5 shows the reconstruction results on the validation images from ImageNet-1K validation set. Based on the reconstruction results on the training images, we observed the overfitting model (*i.e.* SwinV2-L pre-trained on ImageNet1K(10%)) is more like to "remembering" the masked regions, while the non-overfitting model (*i.e.* SwinV2-L pre-trained on ImageNet1K(100%)) is more like "reasoning" the masked regions. For example, the results on the left of the first row in Figure. 4 shows that the overfitting model

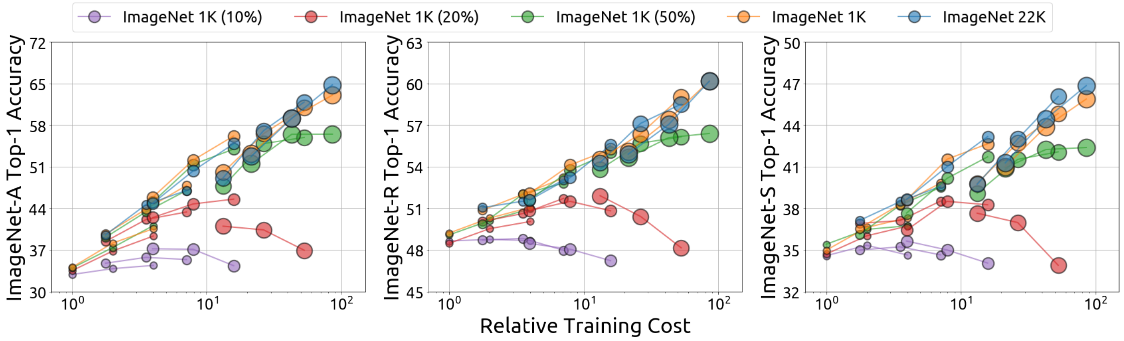

Figure 1. The curves of Top-1 accuracy on ImageNet-A, ImageNet-R and ImageNet-Sketch of different model sizes, data sizes and training lengths, w.r.t. the relative training cost. We set the training cost of SwinV2-S for 125K iterations as the value of 1. Bigger circles indicate larger models. *Best viewed in color.*

"successfully" completes the black hair of the dog, while the non-overfitting model complete the same region in white, since it is a white dog based on the seen regions. In addition, we further observed that the overfitting model appears to lack the "reasoning" ability and overall poorer quality on the validated images compared to the non-overfitting model. For example, the results on the left of the first row in Figure. 5 shows the overfitting model failed to completed the eyes of dog.

D. Cross-Domain Transfer and Robustness Check

To further validate the transfer ability and robustness of different models for different image domains, we conduct more experiments on ImageNet-A [3], ImageNet-R [2] and ImageNet-Sketch [4]. We used models fine-tuned on ImageNet-1K and validate them directly on these datasets. The results shown in Figure. 1 indicates that our conclusions are consistent across different image domains.

References

[1] Yuxin Fang, Li Dong, Hangbo Bao, Xinggang Wang, and Furu Wei. Corrupted image modeling for self-supervised visual pre-training. *arXiv preprint arXiv:2202.03382*, 2022.

[2] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8340–8349, 2021.

[3] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15262–15271, 2021.

[4] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. *Advances in Neural Information Processing Systems*, 32, 2019.

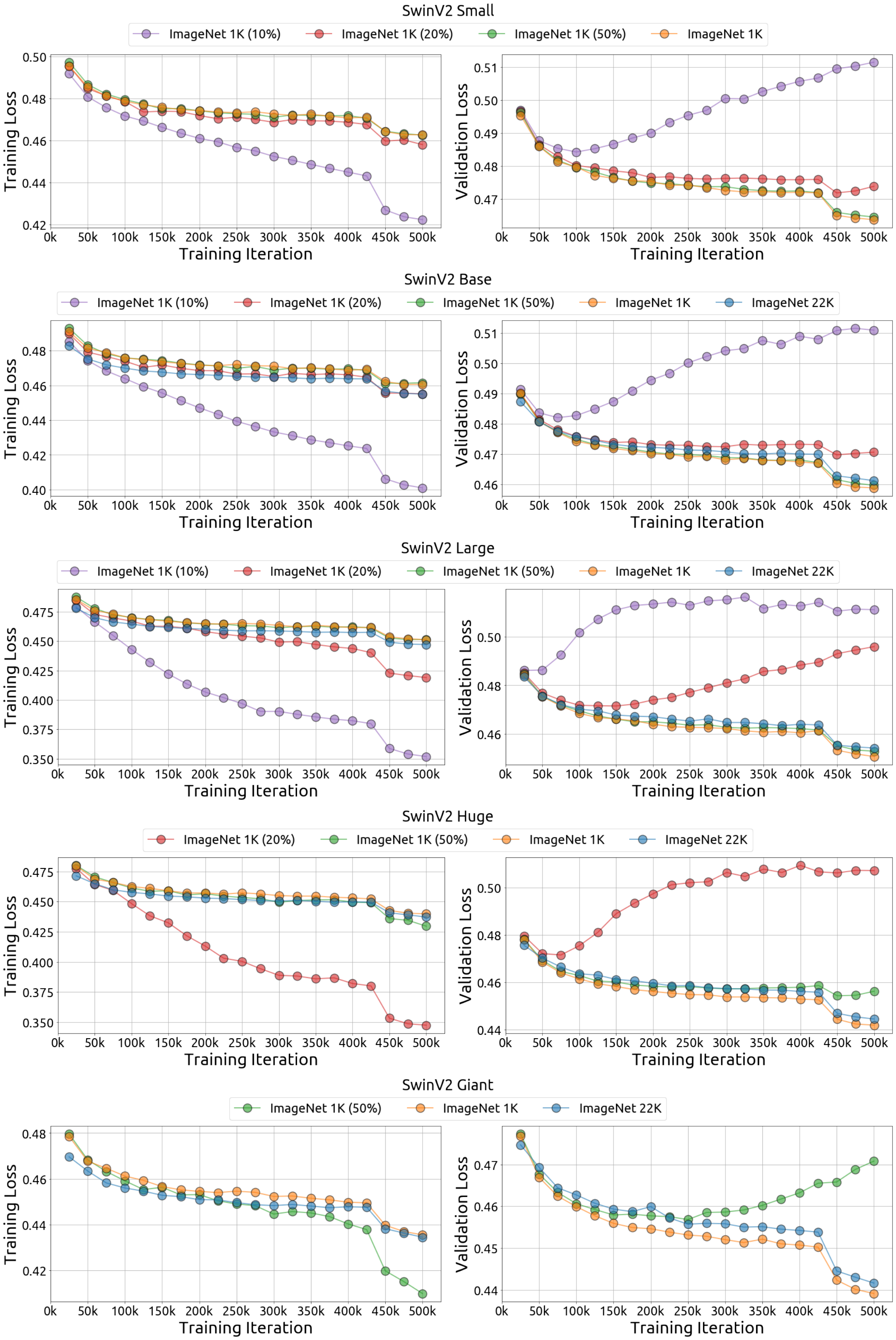

Figure 2. Each row presents the training and the validation loss curves for training with the same model (e.g., SwinV2 giant at the last row) but different datasets. The training loss is computed on its corresponding training dataset, and the validation loss is computed on the ImageNet-1K validation set. *Best viewed in color.*

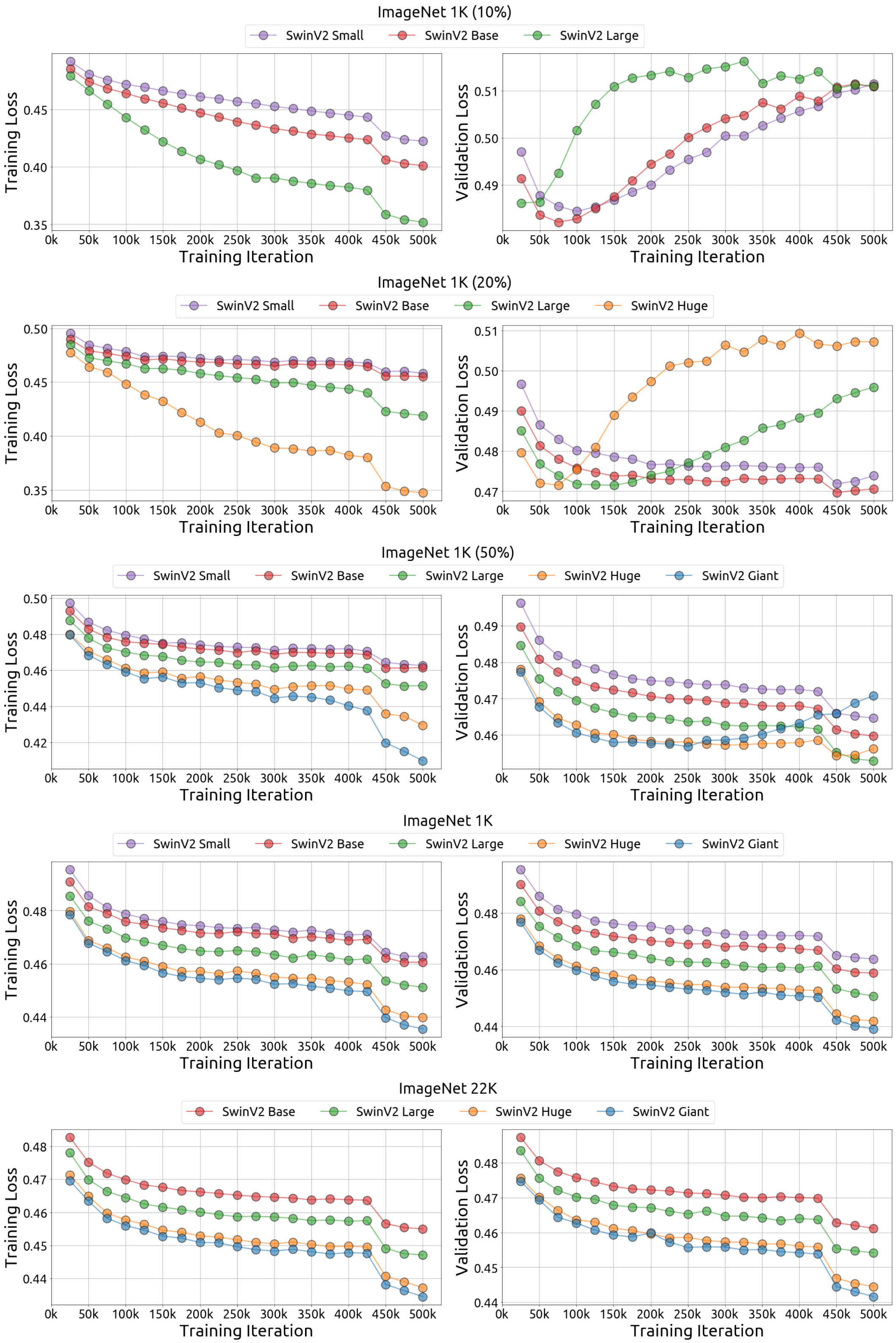

Figure 3. Each row presents the training and the validation loss curves for training with the same dataset (e.g., ImageNet22K at the last row) but different models. The training loss is computed on its corresponding training dataset, and the validation loss is computed on the ImageNet-1K validation set. *Best viewed in color.*

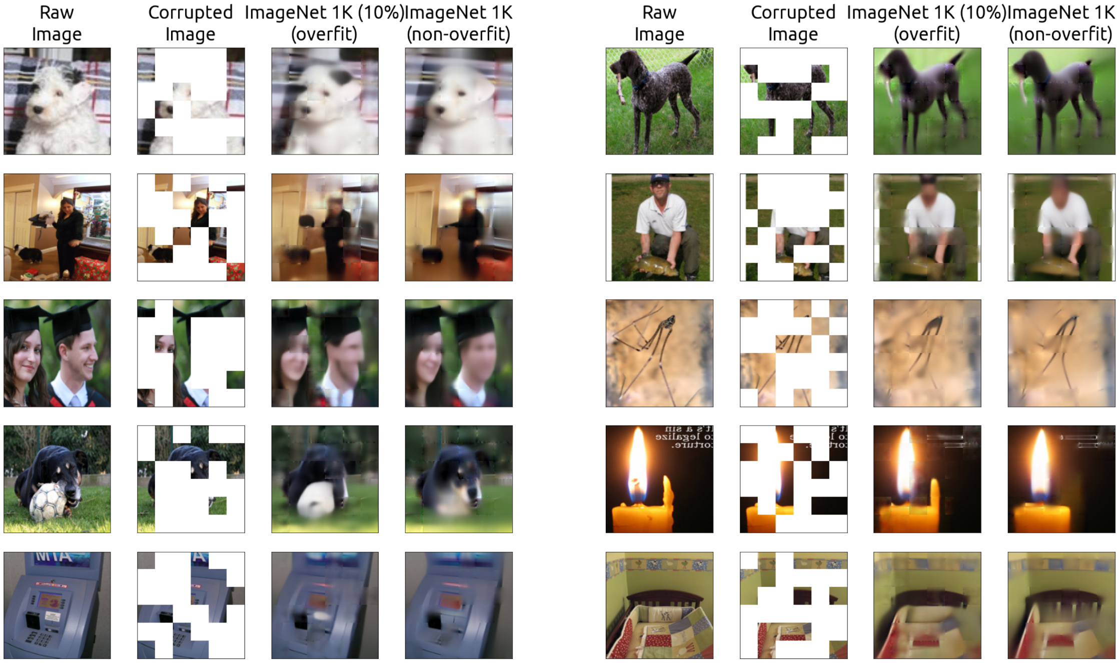
Raw Image
Corrupted Image
ImageNet 1K (10%) (overfit)
ImageNet 1K (non-overfit)
Raw Image
Corrupted Image
ImageNet 1K (10%) (overfit)
ImageNet 1K (non-overfit)

Figure 4. We visualize the reconstruction results of overfitting model (SwinV2-L pre-trained on ImageNet-1K(10%)) and non-overfitting model (SwinV2-L pre-trained on ImageNet-1K(100%)) on the training images from **ImageNet-1K(10%)** dataset, which are jointly contained by the training set of two models. Each group contains 4 images from left to right are: the original image, the corrupted images, reconstructed image of overfitting model, and reconstructed image of non-overfitting model.

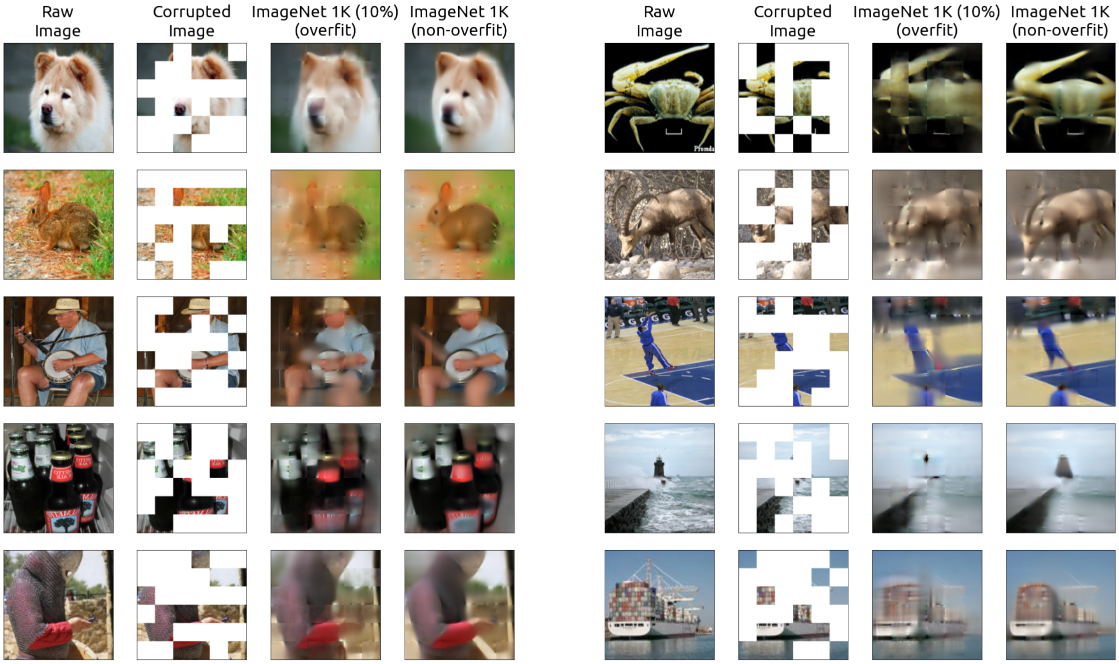
Raw Image
Corrupted Image
ImageNet 1K (10%) (overfit)
ImageNet 1K (non-overfit)
Raw Image
Corrupted Image
ImageNet 1K (10%) (overfit)
ImageNet 1K (non-overfit)

Figure 5. We visualize the reconstruction results of overfitting model (SwinV2-L pre-trained on ImageNet-1K(10%)) and non-overfitting model (SwinV2-L pre-trained on ImageNet-1K(100%)) on the validation images from **ImageNet-1K validation set**. Each group contains 4 images from left to right are: the original image, the corrupted images, reconstructed image of overfitting model, and reconstructed image of non-overfitting model.