Supplementary For: Local Masked Reconstruction for Efficient Self-Supervised Learning on High-resolution Images

Jun Chen[1] *, Faizan Farooq Khan [1] *, Ming Hu[2], Ammar Sherif[3],
Zongyuan Ge[2], Boyang Li[4], Mohamed Elhoseiny[1]
{jun.chen, faizan.khan,mohamed.elhoseiny}@kaust.edu.sa
{ming.hu, zongyuan.ge}@monash.edu
{libo0001@gmail.com, asherif@nu.edu.eg}
[1]King Abdullah University of Science and Technology
[2]Monash University [3]Nile University [4]Nanyang Technological University

1. Experimental Details

1.1. ImageNet Experiments

We launch our experiments by following the MAE (4) settings. The major implementation difference is the model implementation and positional encoding. We apply a 12-layer ViT (3) as a backbone. On top of the ViT, LoMaR adds an MLP layer for the linear projection. We also employ the relative positional encoding (9) in our LoMaR to model the relative positional relation among the patches within the sampled local window.

Pretraining. We apply the batch size 4096 by default during the pretraining. We do not perform any data augmentation strategies. The base learning rate is 1.5e-4. We use AdamW (8) to optimize the model parameters with a weight decay of 0.05. The cosine decay (7) is applied to schedule the learning rate changes. We only apply the RandomResizedCrop augmentation strategy. The warm-up epoch is 40 for pretraining 1,600 epochs, 20 for pretraining 800 epochs, and 10 for pretraining 400 epochs.

Finetuning. We use the pre-trained visual encoder and add another classification head during the finetuning stage. The base learning rate is 1e-3. We apply the adamW (8) optimizer and cosine decay (7) learning rate scheduler in our implementation. We finetune the models for 100 epochs, the same as MAE. The batch size is 1024. The warm-up epoch is 5. The mixup (10) rate is 0.8. We also aggregate the features from all the image patches to generate the whole image representation through average pooling.

1.2. COCO Object Detection Experiments

We follow ViTDet (5) and ViTAE (11) experimental settings and replace the original MAE model with our pre-trained LoMaR. We also integrate our relative positional encoding into their model. For the image resolution of 224×224, we apply our pre-trained LoMaR (4 sampled windows per image + 1,600 pretraining epochs). For the image resolution of 384×384, we apply our pre-trained LoMaR (9 sampled windows per image + 1,600 pretraining epochs). The image patch size is 16×16. We train 25 epochs with a batch size of 64. The input image size is 1024×1024.

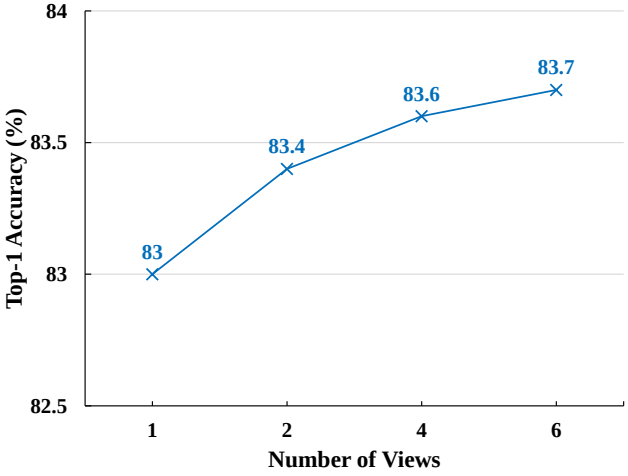

Figure 1. LoMaR performance over different number of views

2. More experimental results

2.1. Ablations on Different Number of Views.

To explore the effect of different numbers of views on the LoMaR pretraining, we additionally sample 1, 2, and 6 views per image. We follow our previous experimental setting to pretrain the model for 400 epochs and finetune the model on the ImageNet-1K dataset. The results can be found in Fig 1.

Method	Absolute Position	Relative Position
MAE	82.5	83.1
LoMaR	**83.1**	**83.5**

Table 1. The results of adding relative positional encoding under local masked reconstruction

RPE vs. APE. Relative positional encoding (RPE) has been widely used in previous works, including BEiT (1). In our LoMaR, RPE can enable our local masked reconstruction with the translation-invariant property, meaning that features of the same object under different regions would not be influenced by the different absolute positional encoding. This is especially more important for our local masked reconstruction compared to the global one. We also employ the RPE (9) in LoMaR. We observe that it can bring 0.4 top-1 accuracy gain from 83.1 to 83.5. Therefore, we set RPE as our default setting for LoMaR.

3. More reconstruction results

We sample more images from ImageNet (2) and MS COCO (6) and perform our local masked reconstruction. The visualization results are shown in the Fig. 2 and 3.

References

[1] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. BEit: BERT pre-training of image transformers. In *International Conference on Learning Representations*, 2022. 2

[2] Jia Deng, Wei Dong, R Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009. 2

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1

[4] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. *arXiv preprint arXiv:2111.06377*, 2021. 1

[5] Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. *arXiv preprint arXiv:2203.16527*, 2022. 1

[6] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 2

[7] Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*, 2017. 1

[8] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2018. 1

[9] Kan Wu, Houwen Peng, Minghao Chen, Jianlong Fu, and Hongyang Chao. Rethinking and improving relative position encoding for vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10033–10041, 2021. 1, 2

[10] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018. 1

[11] Qiming Zhang, Yufei Xu, Jing Zhang, and Dacheng Tao. Vitaev2: Vision transformer advanced by exploring inductive bias for image recognition and beyond. *arXiv preprint arXiv:2202.10108*, 2022. 1

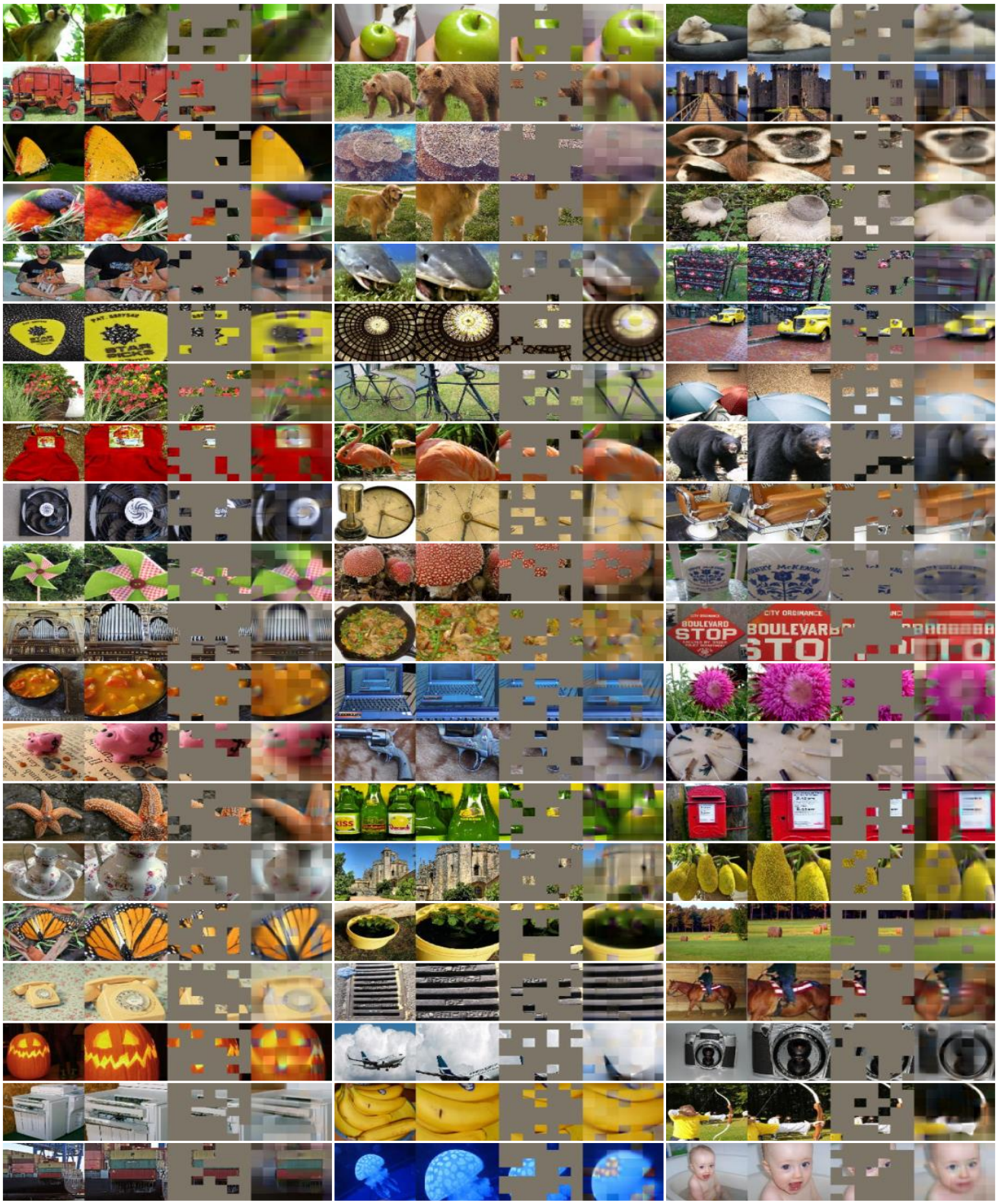

Figure 2. More reconstruction examples of our pretrained model on ImageNet validation images. The masking ratio is 80%. For each image reconstruction figure, we split them into 4 parts: 1) the left-most is the original image. 2) the second-left is the sampled window (7×7 patches). 3) The second-right is the masked image. 4) The right-most is our reconstructed image

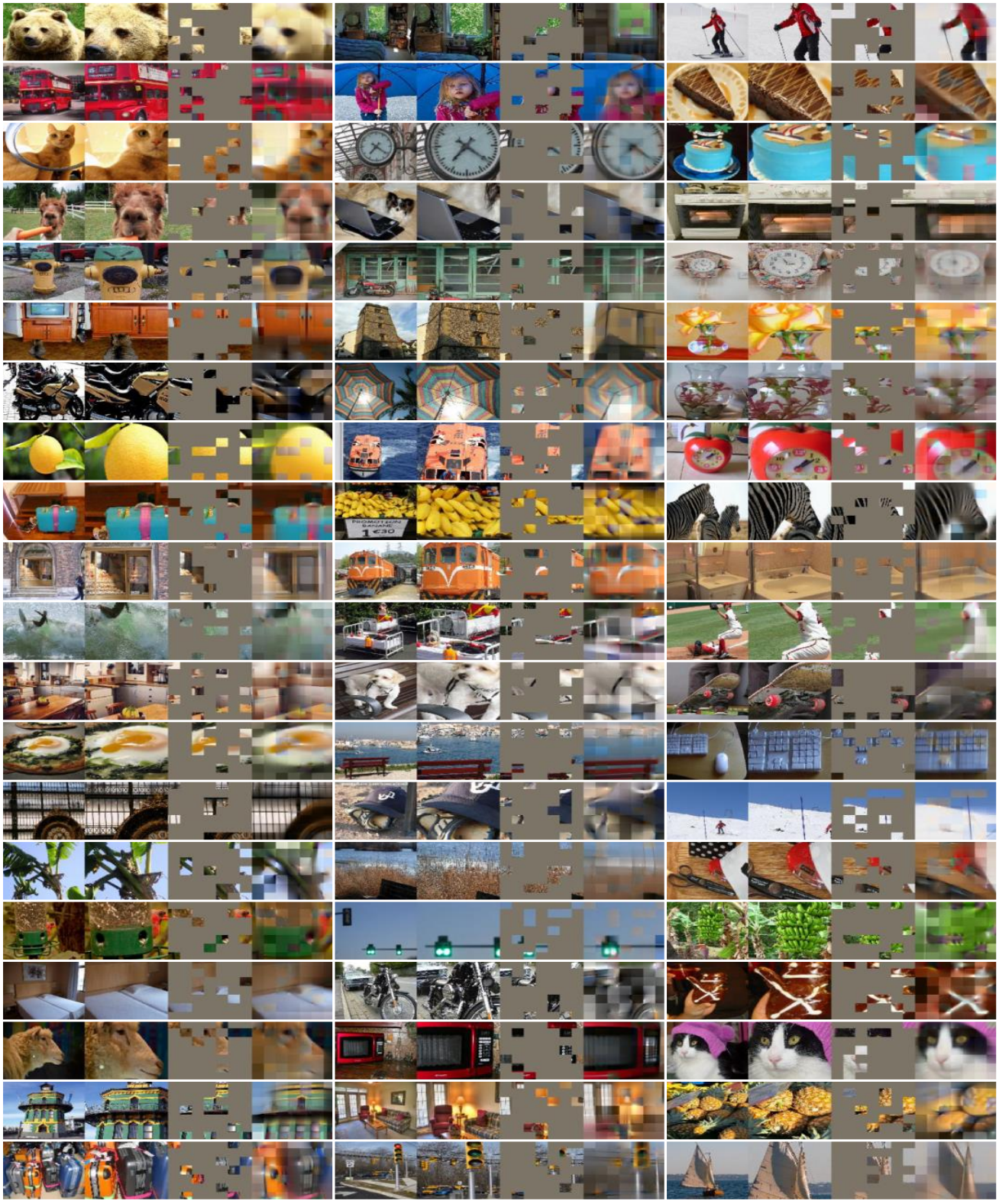

Figure 3. More reconstruction examples of our pretrained model on COCO validation images. The masking ratio is 80%. For each image reconstruction figure, we split them into 4 parts: 1) the left-most is the original image. 2) the second-left is the sampled window (7×7 patches). 3) The second-right is the masked image. 4) The right-most is our reconstructed image