Improving Adversarial Robustness of Masked Autoencoders via Test-time Frequency-domain Prompting Supplementary Materials

Qidong Huang[1] Xiaoyi Dong[2] Dongdong Chen[3] Yinpeng Chen[3] Lu Yuan[3]
Gang Hua[4] Weiming Zhang[1,*] Nenghai Yu[1]
[1]University of Science and Technology of China [2]Shanghai AI Lab
[3]Microsoft Reaserch [4]Wormpex AI Research
{hqd0037@mail., zhangwm@, ynh@}ustc.edu.cn dongxiaoyi@pjlab.org.cn
{cddlyf@, ganghua@}gmail.com {yiche@, luyuan@}microsoft.com

A. Limitations and Social Impacts

For limitations, the first one might be the scope of this work. Our paper reveals the weakness of masked autoencoders (MAE) [15] on adversarial robustness and only strives to mitigate this weakness on image classification problem, of which the research scope seems relatively small. However, we should argue that MAE is far more than just an advanced solution for vision pretraining, it has developed a lot of vision backbone variants [4, 12, 17, 9, 6, 1] and inspires a variety of fields such as video learning [24, 11, 23], 3D point cloud learning [21, 27, 14], multi-modal learning [2, 13], audio learning [26], graph learning [16] or even robotics [25, 22]. The potentials and impacts brought by MAE is tremendous, enough to shock the whole community of deep learning or machine learning. Moreover, pretraining-finetuning has become a more effective way for many downstream vision tasks. Accordingly, an intensive study and improvement for some particular property of MAE cannot be blamed for being too small in scope. By contrast, it should be concerned and encouraged as the result of the popularity of large pretrained models in both industry and academia.

For social impacts, we believe our work can further facilitate the robustness research on large pretrained models or paradigms. With the popularity of the pretraining-finetuning in downstream domains, our work is meaningful to the security/robustness issues existed in such paradigms.

B. Details about Integrating Perceptual Loss in MAE Pretraining

With the analysis given in Sec. 3.2 in our manuscript, it is clear that MAE relies more on medium-/high-frequency signals of input images due to only predicting raw pixels. In the light of BEiT [3] and PeCo [8], a natural idea for improving MAE's ability on adversarial robustness is involving some semantic context in reconstruction rather than only moving towards raw image pixels. To realize this, we leverage a pretrained ViT-Tiny to compute the multi-layer feature distance between MAE decoded image and original image tokens in the masked region, which is so-called perceptual loss [18]. Formally, Let $f_l(x)$ be the normalized feature of l-th transformer block of the pretrained ViT-tiny model. Then for the original image x and the reconstructed image $\hat{x}$, we can formulate the perceptual loss as:

$$L_{perc} = \sum_{l \in \{3,6,9,12\}} \parallel f_l(x) - f_l(\hat{x}) \parallel_2^2, \tag{1}$$

Here we empirically use the feature of $3^{rd}, 6^{th}, 9^{th}, 12^{th}$ transformer block, and the final perceptual loss is the sum of feature loss of these four blocks. The final pretraining objective loss is the sum of the original pixel-level mean square error loss L_{pix} and perceptual loss L_{perc}:

$$L_{total} = L_{pix} + \lambda * L_{perc}. \tag{2}$$

where λ is set as 1 by default. With this simple design, MAE can reduce its dependence on high-frequency signal of images while inheriting original pixel reconstruction loss to maintain the great learning ability. Note that, this improvement only brings limited extra training budget since we just apply a ViT-Tiny model for perceptual loss computation, which is efficient enough during forward or backward compared to the total training time.

To validate the effectiveness of improved MAE that is equipped with our multi-layer perceptual loss, we test its adversarial robustness under the same setting as given in Table 3 of our manuscript. According to the quantitative results showcased, the involvement of perceptual loss can

*Corresponding author.

Algorithm 1: Adversarial BERT Pretraining

Input: clean image set $\boldsymbol{X}$; transformer encoder $\mathbf{E}$; auxiliary prediction head $\mathbf{h}$; pixel reconstruction loss L_{pix}

Output: Parameters θ of $\mathbf{E}$ and $\mathbf{h}$;

for *each sampled mini-batch* $\{\mathbf{x}\} \in \boldsymbol{X}$ **do**

- Generate random masks $\{\mathbf{m}\}$ for $\{\mathbf{x}\}$
- Forward masked mini-batch $\{\mathbf{x_m}\}$ to get reconstructed results $\{\hat{\mathbf{x}}\}$:
  $\hat{\mathbf{x}} = \mathbf{h}(\mathbf{E}(\mathbf{x_m}))$
- Generate the adversarial mini-batch $\{\mathbf{x}_{adv}\}$ with stand untargeted ℓ_∞ PGD attack:

for *t=1,$\cdots$,T* **do**

$\mathbf{x}^t_{adv} = \mathbf{x}^{t-1}_{adv} + \epsilon \nabla_{\mathbf{x}^{t-1}_{adv}} [\mathbf{m} \cdot L_{pix}(\mathbf{x}, \hat{\mathbf{x}}^{t-1}_{adv})]$

end

- Feed the adversarial mini-batch $\{\mathbf{x}_{adv}\}$ into pretraining by minimizing:
  $L_{pix}(\mathbf{x}, \hat{\mathbf{x}}_{adv})$

end

comprehensively boost the ability of MAE. Both clean classification accuracy and adversarial robustness are improved to a nearly satisfying level.

C. Details about Adversarial BERT pretraining

Besides the first baseline, we also consider another intuitive baseline, i.e., introduce adversarial learning into MAE pretraining to improve its robustness. Similar to conventional adversarial learning, our intuition is to force the transformer model to learn on corrupted images that are adversarially generated by gradient-based attack in an online way. The difference is that we want to increase the robustness of pretraining rather than the supervised trained model. Subsequently, we do not have any label information during BERT pretraining to guide the adversarial sample generation.

As shown in Algorithm 1, when integrating adversarial BERT pretraining with MAE, we follow the original MAE pretraining procedure and add the online generated adversarial samples into training alternatively. Since MAE encoder only operates on unmasked tokens, we only add adversarial perturbations onto the unmasked regions. Following untargeted standard ℓ_∞-norm PGD attack, we iteratively generate adversarial masked images by maximizing the MAE pixel reconstruction loss, *i.e.*,

$$\mathbf{x}^t_{adv} = \mathbf{x}^{t-1}_{adv} + \epsilon \nabla_{\mathbf{x}^{t-1}_{adv}} [\mathbf{m} \cdot L_{pix}(\mathbf{x}, \hat{\mathbf{x}}^{t-1}_{adv})], \quad t \in [1, T], \tag{3}$$

where ϵ denotes the attack step size, the superscript$\hat{}$ denotes the reconstructed image, $\mathbf{m}$ is the generated random mask for BERT pretraining, and T is the total iteration number and set as 4 by default. After generating the online adversarial samples $\mathbf{x}_{adv}$, we will add them into MAE pretraining as hard samples and predict the groundtruth raw pixel values from clean samples in an adversarial way.

$$\min_\theta L_{pix}(\mathbf{x}, \hat{\mathbf{x}}_{adv}). \tag{4}$$

Likewise, we conduct the robustness evaluation of MAE over the same setting after equipping with adversarial BERT pretraining. From the results shown in Table 3 of our manuscript, we surprisingly find that both clean classification accuracy and adversarial robustness are boosted. It is a counterintuitive phenomenon, since adversarial training is usually a trade-off game to balance the clean performance and adversarial robustness with proper harness, *i.e.*, one of them can be boosted while the other one degrades. Especially, we find the robustness improvement on C&W and Auto-Attack are not that significant, which may be because C&W and Auto-Attack share different adversarial perturbation generation mechanisms from PGD. This adversarial BERT pretraining gives a successful try to comprehensively improve MAE's ability and will inspire more research works in exploring better transferable adversarial BERT pretraining in this direction.

D. Robustness Evaluation on ℓ_2-Norm Attack

We provide the adversarial robustness results to ℓ_2-norm attack for ViTs that are equipped with different training methods, including the supervised baseline, MoCo v3 [5], BEiT [3], PeCo [8] and MAE [15]. Specifically, we implement the ℓ_2 version of PGD [20], BIM[19], MIM [10] and AA[7] attacks for the verification. For PGD, BIM and MIM, we set the budget as $\epsilon = \varepsilon * \sqrt{C \times H \times W}$, where C, H, W denotes the dimensions of the input image. For AA, we set the budget as η. As the results given in Table 1, MAE's robustness degradation is also significant, further demonstrating the interesting observation raised in our manuscript. **Overall, MAE has obviously degradation on adversarial robustness when compared with other vision BERT pretraining methods.**

E. Layer Correlation

We further provide the layer correlation results for different vision training methods. To better understand the similarity of learned representations (*i.e.*, layer outputs) of these training methods, we tested both 1) the CKA results between different layers of themselves and 2) the CKA results across different layers from different trained ViTs.

From Figure 1, we can easily find that **vision BERT pretraining methods enable the ViT model having the stronger correlation between lower layers and higher layers like previous pretraining methods**, while the ViT trained in supervision shows much worse layer correlation by contrast.

Method	Clean Acc (%)	Robust Acc (%) ($\varepsilon = 0.001, \eta = 0.2$)				Robust Acc (%) ($\epsilon = 0.005, \eta = 0.3$)			
		PGD[20]	MIM[10]	BIM[19]	AA[7]	PGD[20]	MIM[10]	BIM[19]	AA[7]
Supervised	81.8	58.7	39.3	55.8	13.7	0.2	1.1	20.8	4.3
MoCo v3 [5]	83.1	61.1	46.4	59.9	20.9	0.2	1.7	21.5	7.8
BEiT [3]	83.2	64.8	40.8	53.9	19.2	4.6	0.7	15.8	6.8
PeCo [8]	84.5	64.7	41.5	56.4	15.9	3.8	1.2	21.8	4.9
MAE [15]	83.6	**58.4**	**28.9**	**46.3**	**5.9**	**0.1**	**0.3**	**11.6**	**1.3**

Table 1. Comparison of robustness to various ℓ_2-norm adversarial attacks among different BERT training methods. All methods are equally implemented on ViT-Base backbone, and MAE shows significantly worse robustness than other methods. Here, "Clean Acc" and "Robust Acc" mean the classification accuracy on clean samples and adversarial samples respectively. For PGD, BIM and MIM, we set the budget as $\epsilon = \varepsilon * \sqrt{C \times H \times W}$, where C, H, W denotes the dimensions of the input image. For AA, we set the budget as η.

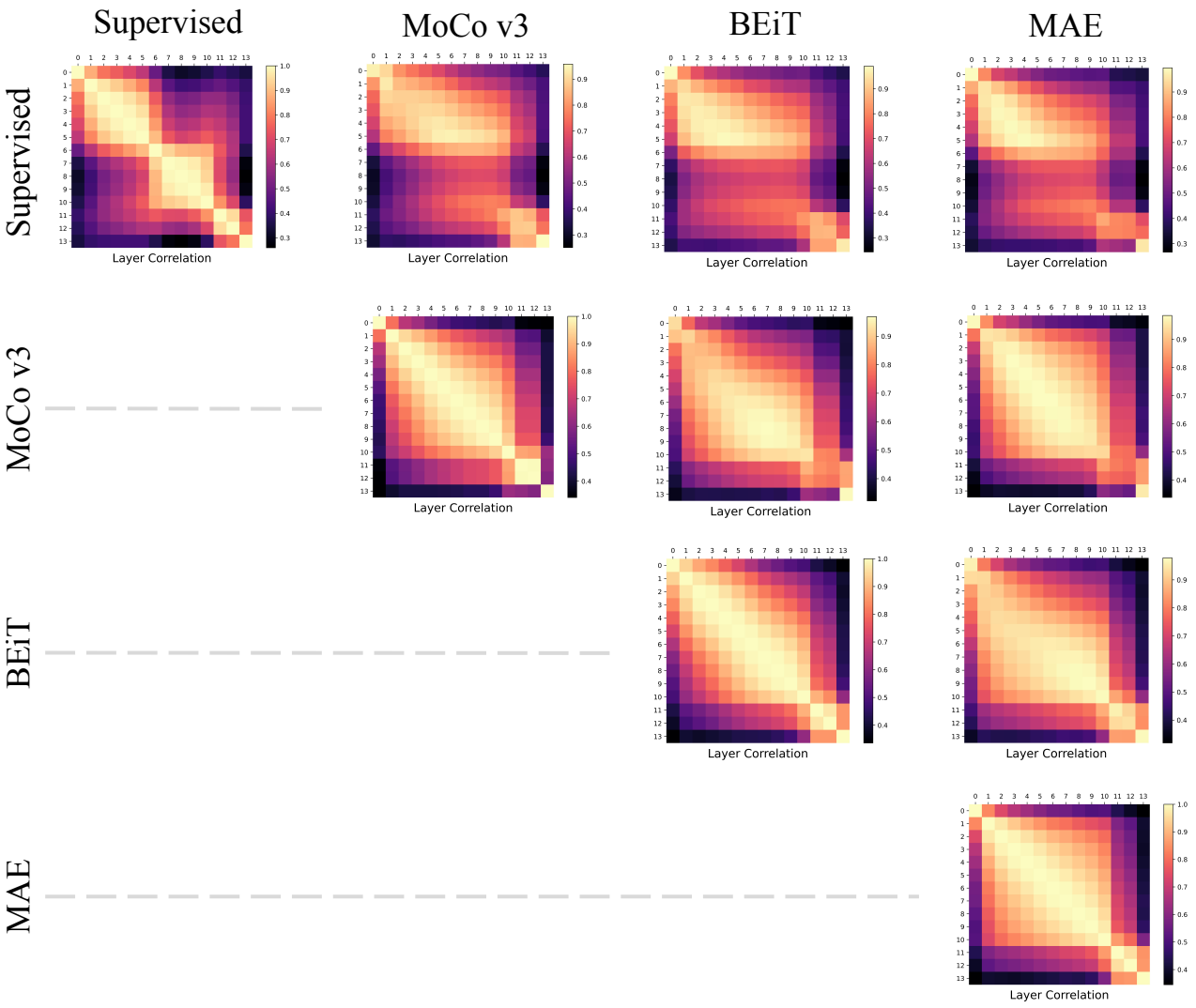

Figure 1. Centered kernel alignment (CKA) similarity calculated on the layer outputs of ViTs that are equipped with different training methods, where "1-12" denotes the different ViT block layer and the layer index "13" means the normalized activation. The brighter lattice denotes the larger CKA similarity, while the darker lattice denotes the weaker CKA similarity.

Besides, it can be observed that **different methods share the similar shallow layer representations**. The reason might be that, as we speculated, they follow the similar way to extract the local information from input images before the feature fusion in the higher layers.

References

[1] Mahmoud Assran, Mathilde Caron, Ishan Misra, Piotr Bojanowski, Florian Bordes, Pascal Vincent, Armand Joulin, Mike Rabbat, and Nicolas Ballas. Masked siamese networks for label-efficient learning. In *ECCV*, 2022. 1

[2] Roman Bachmann, David Mizrahi, Andrei Atanov, and Amir Zamir. Multimae: Multi-modal multi-task masked autoencoders. In *ECCV*, 2022. 1

[3] Hangbo Bao, Li Dong, and Furu Wei. Beit: BERT pre-training of image transformers. *CoRR*, abs/2106.08254, 2021. 1, 2, 3

[4] Xiaokang Chen, Mingyu Ding, Xiaodi Wang, Ying Xin, Shentong Mo, Yunhao Wang, Shumin Han, Ping Luo, Gang Zeng, and Jingdong Wang. Context autoencoder for self-supervised representation learning. *arXiv preprint arXiv:2202.03026*, 2022. 1

[5] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *ICCV*, 2021. 2, 3

[6] Yabo Chen, Yuchen Liu, Dongsheng Jiang, Xiaopeng Zhang, Wenrui Dai, Hongkai Xiong, and Qi Tian. Sdae: Self-distillated masked autoencoder. In *ECCV*, 2022. 1

[7] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *ICML*, 2020. 2, 3

[8] Xiaoyi Dong, Jianmin Bao, Ting Zhang, Dongdong Chen, Weiming Zhang, Lu Yuan, Dong Chen, Fang Wen, and Nenghai Yu. Peco: Perceptual codebook for BERT pre-training of vision transformers. *CoRR*, abs/2111.12710, 2021. 1, 2, 3

[9] Xiaoyi Dong, Jianmin Bao, Ting Zhang, Dongdong Chen, Weiming Zhang, Lu Yuan, Dong Chen, Fang Wen, and Nenghai Yu. Bootstrapped masked autoencoders for vision bert pretraining. In *ECCV*, 2022. 1

[10] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *CVPR*, 2018. 2, 3

[11] Christoph Feichtenhofer, Haoqi Fan, Yanghao Li, and Kaiming He. Masked autoencoders as spatiotemporal learners. *arXiv preprint arXiv:2205.09113*, 2022. 1

[12] Peng Gao, Teli Ma, Hongsheng Li, Jifeng Dai, and Yu Qiao. Convmae: Masked convolution meets masked autoencoders. *arXiv preprint arXiv:2205.03892*, 2022. 1

[13] Xinyang Geng, Hao Liu, Lisa Lee, Dale Schuurams, Sergey Levine, and Pieter Abbeel. Multimodal masked autoencoders learn transferable representations. *arXiv preprint arXiv:2205.14204*, 2022. 1

[14] Ziyu Guo, Xianzhi Li, and Pheng Ann Heng. Joint-mae: 2d-3d joint masked autoencoders for 3d point cloud pre-training. *arXiv preprint arXiv:2302.14007*, 2023. 1

[15] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross B. Girshick. Masked autoencoders are scalable vision learners. *CoRR*, abs/2111.06377, 2021. 1, 2, 3

[16] Zhenyu Hou, Xiao Liu, Yuxiao Dong, Chunjie Wang, Jie Tang, et al. Graphmae: Self-supervised masked graph autoencoders. *arXiv preprint arXiv:2205.10803*, 2022. 1

[17] Zhicheng Huang, Xiaojie Jin, Chengze Lu, Qibin Hou, Ming-Ming Cheng, Dongmei Fu, Xiaohui Shen, and Jiashi Feng. Contrastive masked autoencoders are stronger vision learners. *arXiv preprint arXiv:2207.13532*, 2022. 1

[18] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016. 1

[19] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236*, 2016. 2, 3

[20] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *ICLR*, 2018. 2, 3

[21] Yatian Pang, Wenxiao Wang, Francis EH Tay, Wei Liu, Yonghong Tian, and Li Yuan. Masked autoencoders for point cloud self-supervised learning. In *ECCV*, 2022. 1

[22] Ilija Radosavovic, Tete Xiao, Stephen James, Pieter Abbeel, Jitendra Malik, and Trevor Darrell. Real-world robot learning with masked visual pre-training. *arXiv preprint arXiv:2210.03109*, 2022. 1

[23] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *arXiv preprint arXiv:2203.12602*, 2022. 1

[24] Rui Wang, Dongdong Chen, Zuxuan Wu, Yinpeng Chen, Xiyang Dai, Mengchen Liu, Yu-Gang Jiang, Luowei Zhou, and Lu Yuan. Bevt: Bert pretraining of video transformers. In *CVPR*, 2022. 1

[25] Tete Xiao, Ilija Radosavovic, Trevor Darrell, and Jitendra Malik. Masked visual pre-training for motor control. *arXiv preprint arXiv:2203.06173*, 2022. 1

[26] Hu Xu, Juncheng Li, Alexei Baevski, Michael Auli, Wojciech Galuba, Florian Metze, Christoph Feichtenhofer, et al. Masked autoencoders that listen. *arXiv preprint arXiv:2207.06405*, 2022. 1

[27] Renrui Zhang, Ziyu Guo, Peng Gao, Rongyao Fang, Bin Zhao, Dong Wang, Yu Qiao, and Hongsheng Li. Point-m2ae: multi-scale masked autoencoders for hierarchical point cloud pre-training. *arXiv preprint arXiv:2205.14401*, 2022. 1