SUPPLEMENTARY MATERIAL

Anonymous authors
Paper under double-blind review

In this supplementary material, we introduce more details about (1) the experimental settings, (2) the pre-training loss curves, and (3) the ablation study of different hyperparametric strategies.

0.1 EXPERIMENTAL SETTINGS

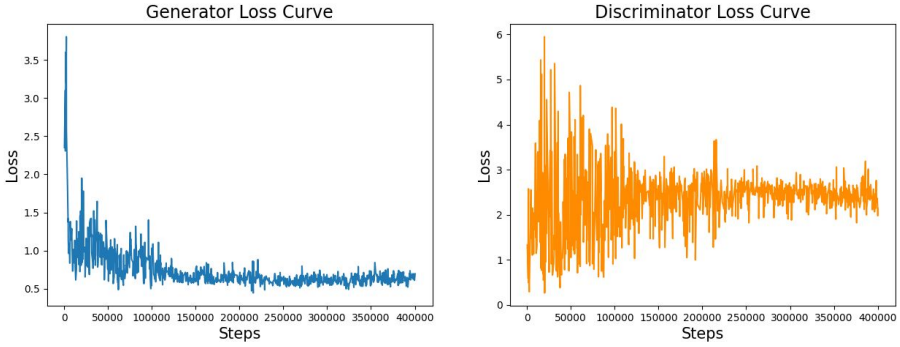
We employ the ViT-B Dosovitskiy et al. (2020) architecture from MAE He et al. (2022) as the generator $\mathcal{G}$ within the framework of Vision ELECTRA $\mathcal{VE}$. For the discriminator $\mathcal{D}$ component of Vision ELECTRA $\mathcal{VE}$, we utilize the ViT-B architecture. Our experimental configuration entails self-supervised pre-training using the ImageNet-1K training dataset Deng et al. (2009). Subsequently, we also have evaluated the learnt representations on downstream tasks such as image classification Lu & Weng (2007), semantic segmentation Guo et al. (2018), and object detection Zou et al. (2023).

During the phase of self-supervised pre-training, the utilization of an AdamW optimizer Kingma & Ba (2014) in conjunction with a cosine learning rate scheduler is employed. The pre-training process spans 50 epochs and is executed on a computing cluster comprising 4x NVIDIA Tesla V100-SXM2 GPUs. Both the generator and discriminator are subjected to akin training hyper-parameters. These parameters encompass a batch size of 80 for each GPU, a base learning rate set at 2e-5, weight decay fixed at 0.05, β_1 at 0.9, β_2 at 0.999, and a warm-up He et al. (2016) period of 10 epochs. A modest data augmentation strategy is implemented, encompassing random resizing cropping with a scale range of [0.67, 1] and an aspect ratio range of [3/4, 4/3], accompanied by random flipping and color normalization procedures.

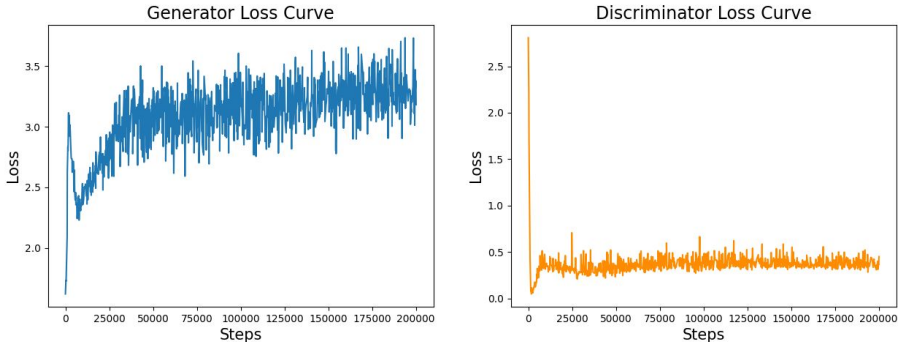
Classification: We initiate the fine-tuning process by utilizing our pre-trained discriminator ViT-B, on the ImageNet-1K image classification task Deng et al. (2009). In the course of fine-tuning, we implement an AdamW optimizer, undertake 100 epochs of training distributed across 4x NVIDIA Tesla V100-SXM2 GPUs, and adopt a cosine learning rate scheduler integrated with a 20-epoch warm-up phase. The hyper-parameters governing the fine-tuning protocol encompass a batch size set at 160, a fundamental learning rate of 1.25e-3, a weight decay rate of 0.05, β_1 at 0.9, β_2 at 0.999, a stochastic depth Huang et al. (2016) ratio of 0.1, and a layer-wise learning rate decay of 0.9. The data augmentation regimen is aligned with that of Bao et al. (2021), encompassing methodologies such as RandAug Cubuk et al. (2020), Mixup Zhang et al. (2017), Cutmix Yun et al. (2019), label smoothing Szegedy et al. (2016), and random erasing Zhong et al. (2020).

Segmentation: We further experiment on ADE20K Zhou et al. (2019) using UperNet Xiao et al. (2018). In fine-tuning, we based on the mae-segmentatio Li (2022) to train our model. The fine-tuning procedure is supported by an AdamW optimizer, encompassing a training duration of 16K iterations distributed across 4x NVIDIA Tesla V100-SXM2 GPUs. The hyper-parameters steering the fine-tuning process comprise a batch size fixed at 2, a foundational learning rate of 1e-4, a weight decay of 0.05, β_1 set to 0.9, β_2 set to 0.999, and an image size configured as 512×512.

Detection: We fine-tune Mask R-CNN He et al. (2017) end-to-end on COCO Lin et al. (2014). The ViT backbone is adapted for use with FPN Lin et al. (2017). During the fine-tuning phase, our approach is built upon the MIMDet Fang et al. (2022) framework. The execution of this procedure is facilitated by the utilization of an AdamW optimizer, spanning a training interval of 100 epochs that is distributed across 4x NVIDIA Tesla V100-SXM2 GPUs. The set of hyper-parameters governing this fine-tuning process encompasses a batch size maintained at 64, an initial learning rate established at 8e-5, a weight decay coefficient of 0.1, β_1 specified as 0.9, β_2 as 0.999, and an image dimension of 768×768.



(a) Pre-training From Scratch Loss Curves



(b) Continual Pre-training Loss Curves

Figure 1: Illustration of Pre-training Loss Curves.

0.2 PRE-TRAINING LOSS CURVES

In this section, we provide the visualization of the loss curves for two different pre-training schemes to further analyze and explain the content of our main manuscript in Section 4.1. We divide our analysis into two parts.

First, in the context of the 'from scratch' scheme (as depicted in Figure 1a), it becomes evident that the $\mathcal{VE}$, which follows the adversarial pre-training manner, is difficult to converge (i.e. drop into the local optimal state) and frequently leads to model collapse (Salimans et al., 2016). This demonstrates that the generator can generate images that are indistinguishable to the discriminator, despite potential deficiencies in realistic or information content (as illustrated in Figure 3 of the main manuscript). Therefore, the model collapse will cause the fluctuations in the training loss of discriminator and can not coverage in a few number of epochs. Consequently, the performance of discriminator demonstrates limited improvement during fine-tuning for downstream tasks.

Secondly, within the framework of the continual pre-training scheme (illustrated in Figure 1b), the utilization of the official pre-trained model results in an advantageous initialization for the discriminator. Therefore, the discriminator can obtain the preliminarily discriminative ability to discern the reconstructed and original images/patches in the early training stages, effectively avoiding the model collapse and coveraging into the stable statue. As the model undergoes optimization, the generator continuously enhance itself performance, producing increasingly realistic images that defy discrimination by the discriminator. Consequently, the image encoder of discriminator can progressively encode more realistic and enhance the hierarchical discrimination throughout representation learning, improving the generalization performance. This, in turn, transfers the superior generalization performance in downstream tasks.

In conclusion, from the Figure 1, we can see that a strong discriminator is necessary, which stabilizes the training process and results in high-quality reconstructed images, leading to better performance.

Figure 2: Examples of different Mask Ratios, e.g., 60%, 85%, 95%.

Strategies	Pre-training strategies	Mask ratio	GAN loss weight	Top-1 acc (%)
Mask Ratio	Official+50ep	60%	0.2	83.01
	Official+50ep	85%	0.2	83.19
	Official+50ep	95%	0.2	82.93
GAN Loss Weight	Official+50ep	75%	0.5	76.71
$\mathcal{VE}$	Official+50ep	75%	0.2	**83.43**

Table 1: Comparisons of Different Mask Ratios and GAN Loss Weights.

0.3 Hyperparametric Strategies

In this section, we have discussed and analyzed the impact of different hyperparameteric strategies on model performance and visual reconstruction results in detail. It contains the following two hyperparameters: 1. Mask ratio, 2. GAN loss weight.

Mask Ratio Due to the mask ratio has an important impact on the quality and diversity of the reconstruction results, therefore, we have delved into the impact of varying mask ratios (e.g. 60%, 85%, 95%) on the quantitative outcomes of the image classification task, as presented in Table 1. Looking at the quantitative results in Table 1, our $\mathcal{VE}$ obtain the best Top-1 acc compared to other mask ratio strategies by using the 75% mask ratio (75% (Our): 83.43 vs. 60%: 83.01 vs. 85%: 83.19 vs. 95%: 82.93). Furthermore, we also present the corresponding reconstruction results for different mask ratios, as illustrated in Figure 2. We can observe that since the 60% mask ratio can make the generator reconstruct the authenticity of images nearly as close to the original, the performance of $\mathcal{VE}$ gradually proximities the 'Official+50ep' effect. Moreover, as the mask ratio increases, it becomes more difficult for the generator to recover more realistic images, which can improve the representation learning of the discriminator. Specifically, for 95% mask ratio, the extremely less visible patches results in a reduced benefit of the generated images for pre-training in the hierarchical discrimination.

During the pre-training of $\mathcal{VE}$ with 95% mask ratio, we also find an interesting observation. Diversity of generated image with low mask ratio: As shown in Figure 3, although most of generated images are visible collapse or invaluable to cause the performance reduction, constrained by the hierarchical discrimination loss, the generator is still capable of the imagination to make generated

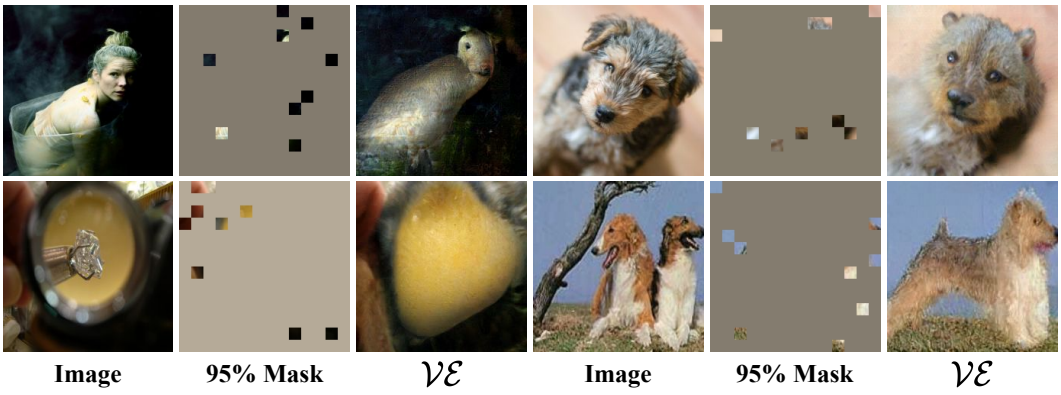


Figure 3: Examples of 95% Mask Ratio.

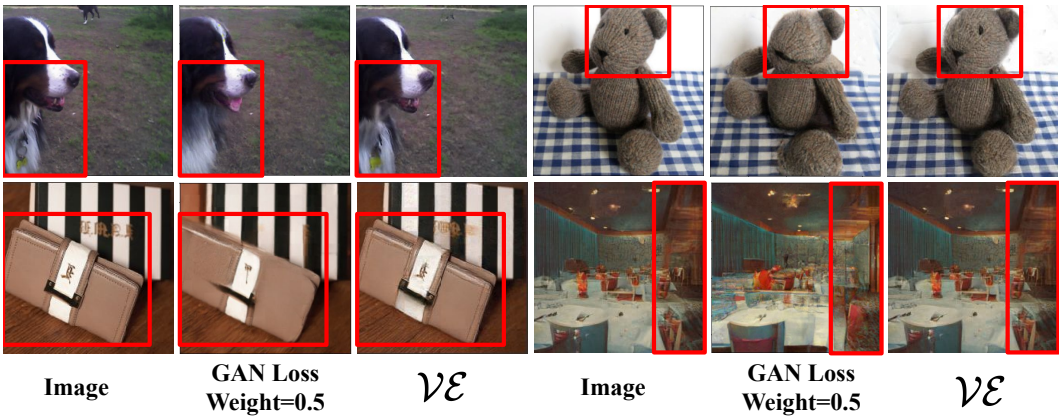


Figure 4: Examples of different GAN Loss Weights. $\mathcal{VE}$ adopts the GAN Loss weight as 0.2 in all experiments.

images to be more reasonable. For the intention of discussion, we believe that future exploration of a generator that generates high-quality images based on a low mask ratio, which is constrained by the hierarchical discrimination loss, can increase the diversity of images and thus further improve the performance of the model.

GAN Loss Weight As the GAN loss significantly contributes to the authenticity of reconstruction results, we engage in a comprehensive discussion and analysis of the influence exerted by varying GAN loss weights (e.g. 0.5). We delve into its ramifications on the enhancement of reconstruction authenticity and, in turn, its repercussions on the generalization performance of discriminator in downstream tasks. The quantitative results are presented in Table 1. As shown in Table 1, we can observe that adopting the larger GAN loss weight will cause an extreme reduction of the improvement fine-tuning the downstream task, since a large weight leads to the generator saturation (low-quality generated images) and unstable training (Goodfellow et al., 2014). Furthermore, we also present the corresponding reconstruction results for with and without GAN loss weights, as illustrated in Figure 4. We highlight the differences among several kinds of images: the original image, the images generated with GAN loss weight=0.5 and the reconstructed images obtained by $\mathcal{VE}$. It is clear to demonstrate that adopting a low but appropriate GAN loss weight can generate the high-quality images to improve the performance of model. In the experiment, we set the GAN loss weight to 0.2.

REFERENCES

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 702–703, 2020.

Jia Deng, Wei Dong, Richard 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*, pp. 248–255. Ieee, 2009.

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.

Yuxin Fang, Shusheng Yang, Shijie Wang, Yixiao Ge, Ying Shan, and Xinggang Wang. Unleashing vanilla vision transformer with masked image modeling for object detection. *arXiv preprint arXiv:2204.02964*, 2022.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

Yanming Guo, Yu Liu, Theodoros Georgiou, and Michael S Lew. A review of semantic segmentation using deep neural networks. *International journal of multimedia information retrieval*, 7:87–93, 2018.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969, 2017.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16000–16009, 2022.

Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q Weinberger. Deep networks with stochastic depth. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pp. 646–661. Springer, 2016.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Xiang Li. Ade20k semantic segmentation with mae. `https://github.com/implus/mae_segmentation`, 2022.

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 *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755. Springer, 2014.

Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2117–2125, 2017.

Dengsheng Lu and Qihao Weng. A survey of image classification methods and techniques for improving classification performance. *International journal of Remote sensing*, 28(5):823–870, 2007.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.

Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 418–434, 2018.

Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6023–6032, 2019.

Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 13001–13008, 2020.

Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127:302–321, 2019.

Zhengxia Zou, Keyan Chen, Zhenwei Shi, Yuhong Guo, and Jieping Ye. Object detection in 20 years: A survey. *Proceedings of the IEEE*, 2023.