

---

# Visualizing the loss landscape of Self-supervised Vision Transformer

---

<sup>1,2</sup>Youngwan Lee <sup>2</sup>Jeffrey Willette <sup>1</sup>Jonghee Kim <sup>2</sup>Sung Ju Hwang

<sup>1</sup>Electronics and Telecommunications Research Institute (ETRI), South Korea

<sup>2</sup>Korea Advanced Institute of Science and Technology (KAIST), South Korea

## Abstract

The Masked autoencoder (MAE) has drawn attention as a representative self-supervised approach for masked image modeling with vision transformers. However, even though MAE shows better generalization capability than fully supervised training from scratch, the reason why has not been explored. In another line of work, the Reconstruction Consistent Masked Auto Encoder (RC-MAE), has been proposed which adopts a self-distillation scheme in the form of an exponential moving average (EMA) teacher into MAE, and it has been shown that the EMA-teacher performs a conditional gradient correction during optimization. To further investigate the reason for better generalization of the self-supervised ViT when trained by MAE (MAE-ViT) and the effect of the gradient correction of RC-MAE from the perspective of optimization, we visualize the loss landscapes of the self-supervised vision transformer by both MAE and RC-MAE and compare them with the supervised ViT (Sup-ViT). Unlike previous loss landscape visualizations of neural networks based on classification task loss, we visualize the loss landscape of ViT by computing pre-training task loss. Through the lens of loss landscapes, we find two interesting observations: (1) MAE-ViT has a smoother and wider overall loss curvature than Sup-ViT. (2) The EMA-teacher allows MAE to widen the region of convexity in both pretraining and linear probing, leading to quicker convergence. To the best of our knowledge, this work is the first to investigate the self-supervised ViT through the lens of the loss landscape.

## 1 Introduction

Due to the scalability and versatility of self-attention mechanisms [35], the Vision Transformer (ViT) [14] has been widely used in the vision domain from image/pixel-level recognition [29, 16, 24, 15] to video applications [17, 34, 3]. In the recent self-supervised learning literature [2, 19, 38, 1, 13, 25], masked image modeling such as the Masked autoencoder [19] (MAE) which utilizes a ViT backbone and predicts masked patches given a set of unmasked patches, has been a staple pre-text task. MAE achieves better generalization performance than a fully supervised ViT from scratch, but, the reasons for the superior generalization have not been widely explored yet.

Meanwhile, Lee et al. [25] have proposed the reconstruction-consistent Masked autoencoder (RC-MAE) by adopting a self-distillation scheme [18, 5] in the form of an exponential moving average (EMA) teacher into the MAE framework. The RC-MAE showed that the EMA-teacher performs conditional gradient corrections during optimization. As a result, RC-MAE showed better generalization performance on downstream tasks and achieved a faster convergence speed than MAE.

In this work, we analyze the reason for better generalization of the self-supervised ViT when trained by MAE and the effect of the gradient correction of RC-MAE from the perspective of optimization. To do this, we visualize loss landscapes [26] of the self-supervised ViTs as shown in Fig. 1 (MAE) and compare them with the supervised ViT trained from scratch. Specifically, we compute the pretraining

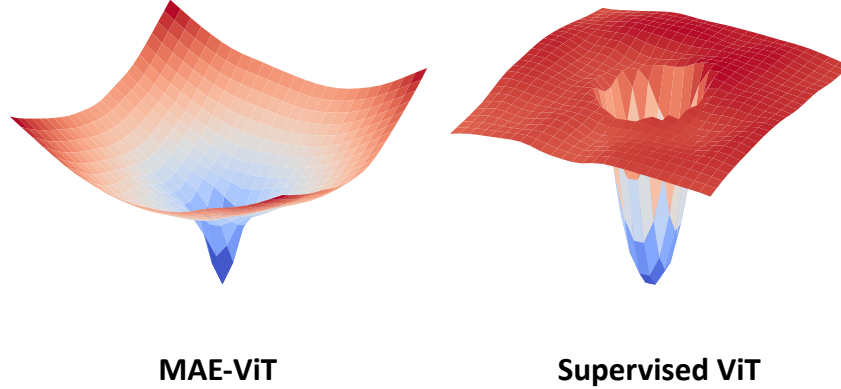


Figure 1: **Comparison between self-supervised vision transformer (ViT) [14] by MAE [19] and supervised ViT.** MAE-ViT is drawn by the pre-training loss (masked patch prediction) and Supervised ViT is obtained by supervised image classification from scratch. We use the ViT-Base model. MAE-ViT converges at a *smoother* and *wider* convex region than the supervised ViT.

losses (e.g., masked patch reconstruction) in MAE [19] and RC-MAE [25] and visualize the loss landscapes by using the filter normalization method [26].

Previous works [22, 23, 21, 32, 6, 10] relating optimization with the loss landscape have demonstrated that a *flatter* region with smaller curvature correlates well with the generalization of neural networks. [10] investigated the loss landscape of a supervised ViT from scratch. However, these works have analyzed convolutional neural networks or vision transformers in a fully supervised learning setting, e.g., the image-classification task with cross-entropy loss. To the best of our knowledge, our work is the first study to visualize the loss geometry of *self-supervised* vision transformers.

Through the lens of the loss landscape, we summarize our findings on the self-supervised ViT as follows:

- Self-supervised ViTs have a *flatter* and *smoother* overall loss curvature and show better generalization properties compared to supervised ViT.
- The addition of an EMA-teacher [25] allows MAE to widen the region of convexity, which leads to faster convergence.

In this work, we qualitatively analyze the optimized loss landscapes of the self-supervised ViT. We leave more conclusive quantitative experiments for future works. We hope that our work provokes further analysis into the self-supervised vision transformer from the perspective of optimization.

## 2 Preliminary: MAE and RC-MAE

**The Masked Autoencoder (MAE)** [19] performs self-supervised learning by randomly masking a large portion of input patches and then using a ViT-based encoder  $f$  and subsequently reconstructing the masked patches with decoder  $h$ , given the encoded visible patches. MAE splits the input image  $X \in \mathbb{R}^{C \times H \times W}$  into  $N$  disjoint patches  $\tilde{X} \in \mathbb{R}^{N \times (P^2 \cdot C)}$  where  $P^2$  represents the area of a patch. MAE then masks a random subset of  $\mathbf{x}_i \in \tilde{X}$ ,  $\forall i \in \mathcal{M}$ , with  $\mathcal{M}$  being the indices of the mask tokens. The visible patches,  $\mathbf{x}_j \in \tilde{X}$ ,  $\forall j \in \mathcal{V}$  (with  $\mathcal{V}$  being the indices of the visible patches) are given to the encoder which produces an encoded vector  $z = f(\{\mathbf{x}_j : j \in \mathcal{V}\})$ . Subsequently, the decoder  $h$  performs a reconstruction of the set of masked patches  $\{\mathbf{x}_i : i \in \mathcal{M}\}$  given the encoded visible patches,  $\hat{Y} = h(z; \{\mathbf{x}_j : j \in \mathcal{M}\})$ , where  $\hat{Y} \in \mathbb{R}^{N \times (P^2 \cdot C)}$ . The loss function  $\mathcal{L}_r$  is then only computed on the masked patches with a mean squared error loss function.

$$\mathcal{L}_r = \frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} \|\tilde{X}_i - \hat{Y}_i\|_2^2. \quad (1)$$

**The Reconstruction Consistent Masked Autoencoder (RC-MAE)** [25] provided analysis into a common technique utilized [18, 5, 38, 1] throughout self-supervised learning, the EMA teacher [33]. An EMA teacher  $T$  is composed of an exponential moving average of previous students  $S$  with

$\alpha \in [0, 1]$  by  $T^{(t)} = \alpha T^{(t-1)} + (1 - \alpha)S^{(t)}$  which can be recursively expanded to  $T^{(t)} = \sum_{i=0}^t \alpha^i (1 - \alpha)S^{t-i}$ . In addition to the MAE’s [19] reconstruction objective  $\mathcal{L}_r$ , the teacher provides a consistency target  $\hat{Y}'$  to the student network (*e.g.*, MAE). Thus, the student network is optimized by the objectives as follows:

$$\frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} (\underbrace{\|\tilde{X}_i - \hat{Y}_i\|^2}_{\text{reconstruction}} + \underbrace{\|\hat{Y}_i - \hat{Y}'\|^2}_{\text{consistency}}), \quad (2)$$

It was revealed that in a simple linear model [25], the teacher acts in essence like a gradient memory which removes previous gradient directions conditionally if the current input  $\mathbf{x}_i$  is similar to a previous input  $\hat{\mathbf{x}}_j$  measured by the dot product. Likewise, when the current inputs are orthogonal to previous inputs, the dot product is 0 and the teacher gives no corrective signal. Thus, the EMA teacher may stabilize training by actively preventing overfitting when it is likely (*e.g.* when there is low input diversity) and allowing the model to learn new knowledge when overfitting is less likely (*e.g.* there is high input diversity).

### 3 Loss landscape

To visualize loss landscapes of vision transformer, we follow the visualization strategy from [26] called filter-wise normalization. Specifically, [26] obtains two random Gaussian direction vectors  $\delta$  and  $\eta$  for each parameter  $\theta$  to visualize the loss surface within the 2D projected space. Note that the direction vectors are normalized to have the same norm as the corresponding parameter  $\theta$ . And then, the loss surface is obtained by evaluating the loss on 2D points along the two directions as follows:

$$f(\alpha, \beta) = L(\theta + \alpha\delta + \beta\eta), \quad (3)$$

where  $L$  is the loss function for a network parameterized by  $\theta$ .  $\alpha$  and  $\beta$  are varied scalar values from -1 to 1 corresponding to the  $x$ -axis and  $y$ -axis in the loss surface, respectively. please refer to [Appendix A](#) for more implementation details.

#### 3.1 Analysis

[26] demonstrates that the *flatness* of minimizers correlates well with generalization due to this visualization method based on filter normalization. In addition, numerous studies [6, 22, 32, 23, 21, 26, 37, 9, 10] have drawn a conclusion that **neural networks tend to generalize better when they converge to a flat region with small curvature and a wide region of convexity**. Since we already confirmed that self-supervised learning methods, MAE [19] and RC-MAE [25], using vision transformers [14] generalize better than purely supervised learning in [Tab. 1](#), in this section, we analyze the reason for the better generalization capability of the self-supervised methods and investigate the effect of the EMA Teachers in RC-MAE from the *optimization* perspective by observing the loss landscapes.

**MAE-ViT vs. Supervised-ViT.** As shown in [Fig. 1](#), the vision transformer [14] supervised from scratch shows a narrower convergence region (*i.e.*, convex region) which is also observed by [10]. In contrast, we can observe that the self-supervised vision transformer by MAE [19] exhibits a much wider convex area of the loss landscape. This demonstrates that the self-supervised methods likely converge under a broader set of initial conditions compared to fully supervised models. We note that this *smoother* and *wider* loss landscape and better downstream task performance (*i.e.*, generalization result) are consistent with the fact that *a wide convex region with a small curvature correlates well with the generalization of neural networks* in the optimization literature [22, 23, 21, 32, 6, 10].

**MAE vs. RC-MAE.** [Fig. 2](#) illustrates loss landscapes of the self-supervised vision transformers by MAE [19] and RC-MAE [25] which result from pre-training and linear probing. For the pre-training as shown in [Fig. 2](#) (left), the top-view and 2D loss contour of the landscapes show that RC-MAE

Table 1: **Downstream tasks results** using ViT-Base model. FN and LN respectively denote end-to-end finetuning and linear probing on ImageNet-1K. bbox/mask AP are results utilizing Mask R-CNN Benchmarking [27]. These results demonstrate that MIM-based pre-training shows better **generalization** capability than fully supervised training from scratch.

Method	FN	LN	bbox AP	mask AP
Supervised	82.3	-	47.9	49.3
MAE [19]	83.4	67.3	50.3	44.9
RC-MAE [25]	<b>83.6</b>	<b>68.4</b>	<b>51.0</b>	<b>45.4</b>

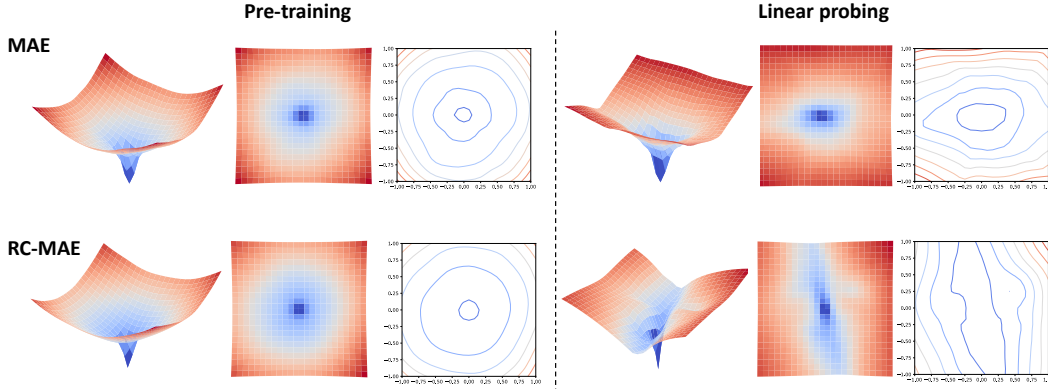


Figure 2: **Comparison between MAE [19] and RC-MAE [25] using pre-training (left) and linear probing (right) weights.** We use the ViT-B which is pre-trained for 1600 epochs with MAE or RC-MAE. While the pre-training computes the masked reconstruction loss through mean-square error loss, the linear probing task computes image classification loss with cross-entropy loss.

converges from a *wider* region of convexity than MAE. Additionally, as shown in Fig. 2 (right), linear probing results of both MAE and RC-MAE have more complex loss curvature than that of the pre-training loss. We speculate that the linear probing task freezing the feature weights and only learning a linear layer is hard to optimize for classifying 1K categories. Similar to the pre-training result, RC-MAE has a wider convex region than MAE. As the only difference between MAE and RC-MAE is the addition of an EMA teacher, this suggests that this wider convex region could be attributed to the effect of the gradient correction by the EMA-teacher in RC-MAE. Furthermore, Lee et al. [25] have demonstrated that the convergence speed of RC-MAE is faster than MAE by comparing reconstruction loss graphs and finetuning accuracies in their paper. These loss landscape comparisons between MAE and RC-MAE together with the experiments in [25] support the better convergence properties of RC-MAE.

## 4 Conclusion and Future works

In this work, we have investigated the reason for the generalization capability of the self-supervised vision transformer and the gradient correction effect of RC-MAE by visualizing the loss landscapes of various self-supervised ViT’s. Through the lens of loss landscapes, we have observed interesting things: (1) Self-supervised vision transformers have a smoother and wider overall loss curvature than fully supervised ViT’s. (2) The self-distillation architecture (*i.e.*, EMA-teacher) allows MAE to widen the region of convexity, accelerating convergence speed. However, there is still room for further exploring the effect of self-supervised learning. **Batch size:**[26] shows that batch size affects the sharpness of the minimizer. Future works could explore the effect of the batch size on MIM pre-training. **Optimizers:** [10] utilize a sharpness-aware optimizer which could have some of the same properties as the EMA teacher in RC-MAE. A direct comparison and analysis could yield some interesting takeaways. **Comparison with other self-supervised methods:** We have analyzed only MIM-based self-supervised methods [19, 25]. However, before the emergence of MIM methods, instance discrimination tasks (*e.g.*, contrastive learning)-based methods [20, 11, 12, 7, 18, 8, 5] had been dominant. Therefore, it would be interesting to investigate and compare the non-MIM methods from the standpoint of loss geometry. **Quantitative analysis:** We have performed only qualitative analyses using loss landscapes. However, [10] quantifying the *average flatness* and the degree of loss *curvature* by calculating the training error under Gaussian perturbations on the model parameter and the dominant Hessian eigenvalue, respectively. Thus, we look forward to future works which may quantitatively compare those metrics of optimization dynamics.

### Acknowledgement

This work was partly supported by Institute of Information & Communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (No. RS-2022-00187238, Development of Large Korean Language Model Technology for Efficient Pre-training (80%) and No. 2020-0-00004, Development of Previsional Intelligence based on Long-term Visual Memory Network (20%)).

## References

- [1] Mahmoud Assran, Mathilde Caron, Ishan Misra, Piotr Bojanowski, Florian Bordes, Pascal Vincent, Armand Joulin, Michael Rabbat, and Nicolas Ballas. Masked siamese networks for label-efficient learning. In *ECCV*, 2022. 1, 2, 7
- [2] Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. In *ICLR*, 2022. 1
- [3] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, 2021. 1
- [4] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *NeurIPS*, 2020. 7
- [5] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021. 1, 2, 4, 7
- [6] Pratik Chaudhari, Anna Choromanska, Stefano Soatto, Yann LeCun, Carlo Baldassi, Christian Borgs, Jennifer Chayes, Levent Sagun, and Riccardo Zecchina. Entropy-SGD: Biasing gradient descent into wide valleys. In *ICLR*, 2017. 2, 3
- [7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020. 4, 7
- [8] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. Big self-supervised models are strong semi-supervised learners. *NeurIPS*, 33:22243–22255, 2020. 4
- [9] Xiangning Chen and Cho-Jui Hsieh. Stabilizing differentiable architecture search via perturbation-based regularization. In *ICML*, 2020. 3
- [10] Xiangning Chen, Cho-Jui Hsieh, and Boqing Gong. When vision transformers outperform resnets without pre-training or strong data augmentations. In *ICLR*, 2022. 2, 3, 4, 7
- [11] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. In *arXiv preprint arXiv:2003.04297*, 2020. 4, 7
- [12] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *ICCV*, 2021. 4, 7
- [13] 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
- [14] 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. In *ICLR*, 2021. 1, 2, 3, 7
- [15] Alaaeldin El-Nouby, Hugo Touvron, Mathilde Caron, Piotr Bojanowski, Matthijs Douze, Armand Joulin, Ivan Laptev, Natalia Neverova, Gabriel Synnaeve, Jakob Verbeek, et al. Xcit: Cross-covariance image transformers. In *NeurIPS*, 2021. 1
- [16] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *ICCV*, 2021. 1
- [17] Christoph Feichtenhofer, Yanghao Li, Kaiming He, et al. Masked autoencoders as spatiotemporal learners. *NeurIPS*, 2022. 1
- [18] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent—a new approach to self-supervised learning. In *NeurIPS*, 2020. 1, 2, 4, 7
- [19] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022. 1, 2, 3, 4, 7
- [20] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020. 4, 7
- [21] Stanisław Jastrzebski, Zachary Kenton, Nicolas Ballas, Asja Fischer, Yoshua Bengio, and Amos Storkey. On the relation between the sharpest directions of DNN loss and the SGD step length. In *ICLR*, 2019. 2, 3

- [22] Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. In *ICLR*, 2017. 2, 3
- [23] Bobby Kleinberg, Yuanzhi Li, and Yang Yuan. An alternative view: When does SGD escape local minima? In *ICML*, 2018. 2, 3
- [24] Youngwan Lee, Jonghee Kim, Jeffrey Willette, and Sung Ju Hwang. Mpvit: Multi-path vision transformer for dense prediction. In *CVPR*, 2022. 1
- [25] Youngwan Lee, Jeffrey Ryan Willette, Jonghee Kim, Juho Lee, and Sung Ju Hwang. Exploring the role of mean teachers in self-supervised masked auto-encoders. In *ICLR*, 2023. 1, 2, 3, 4, 7
- [26] Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets. *NeurIPS*, 2018. 1, 2, 3, 4, 7
- [27] Yanghao Li, Saining Xie, Xinlei Chen, Piotr Dollar, Kaiming He, and Ross Girshick. Benchmarking detection transfer learning with vision transformers. *arXiv preprint arXiv:2111.11429*, 2021. 3, 7
- [28] 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 *ECCV*, 2014. 7
- [29] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021. 1
- [30] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *ICML*, 2019. 7
- [31] Zhiqiang Shen, Zechun Liu, and Eric Xing. Sliced recursive transformer. In *ECCV*, 2022. 7
- [32] Samuel L. Smith and Quoc V. Le. A bayesian perspective on generalization and stochastic gradient descent. In *ICLR*, 2018. 2, 3
- [33] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *NeurIPS*, 2017. 2
- [34] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *NeurIPS*, 2022. 1
- [35] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 1
- [36] Ross Wightman. Pytorch image models. <https://github.com/rwightman/pytorch-image-models>, 2019. 7
- [37] Arber Zela, Thomas Elsken, Tonmoy Saikia, Yassine Marrakchi, Thomas Brox, and Frank Hutter. Understanding and robustifying differentiable architecture search. In *ICLR*, 2020. 3
- [38] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. In *ICLR*, 2022. 1, 2, 7

# Appendix

## A Implementation Details

**Setup.** For reducing computation cost, we compute the loss landscape on the ImageNetV2 [30] validation set which has 10K images. We use the pre-trained weight for the supervised ViT the original authors provided via the `timm` [36] library. As existing loss landscape works [26, 10, 31] visualize neural networks by using a supervised image classification task with cross-entropy loss, we also visualize the loss landscape of the supervised ViT by computing image classification loss.

**Pre-training.** Unlike the supervised ViT using image classification task, MAE [19] and RC-MAE [25] perform a masked patch reconstruction task for pre-training. Since we want to analyze the loss dynamics during the pre-training phase, we perform the filter normalization method [26] with the masked patch reconstruction loss, instead of a classification loss. Thus, we visualize the loss landscapes of MAE and RC-MAE by using Equation (1) and Equation (2), respectively. To do this, firstly, we pre-train both MAE and RC-MAE with vision transformer (ViT [14] base) for 1600 epochs in a 1-node server equipped with 8 GPUs, following the setup of RC-MAE. The visualizations of loss landscapes are then created using the pre-trained weights.

**Linear probing.** Linear probing or fine-tuning on downstream tasks have been used as standard protocols for self-supervised learning [20, 7, 18, 11, 4, 12, 5, 19, 38, 1]. Linear probing freezes the pre-trained weights (*i.e.*, feature) and learns a linear classifier on top of the frozen features in the ImageNet classification task. We assume that this linear probing could also prove to be an insightful task in analyzing the pre-training because the pre-trained features are frozen. To do this, we perform the linear probing evaluation using the pre-trained models for 1600 epochs by MAE and RC-MAE and achieve the results as shown in Tab. 1. We obtain loss landscapes from the linear probing weights of both MAE and RC-MAE by using cross-entropy loss.

**Downstream tasks.** To validate the generalization capability of the vision transformer which is pre-trained by MAE and RC-MAE, we perform downstream tasks including end-to-end ImageNet fine-tuning, object detection, and instance segmentation using Mask R-CNN benchmarking [27] on the COCO [28] dataset. As shown in Tab. 1, MAE and RC-MAE consistently outperform the supervised model in all tasks, demonstrating that self-supervised methods show better generalization than supervised learning. Furthermore, RC-MAE also achieves higher performance than MAE in all downstream tasks.