WHAT DO SELF-SUPERVISED VISION TRANSFORMERS LEARN?

Anonymous authors
Paper under double-blind review

ABSTRACT

We present a comparative study on how and why contrastive learning (CL) and masked image modeling (MIM) differ in their representations and in their performance of downstream tasks. In particular, we demonstrate that self-supervised Vision Transformers (ViTs) have the following properties: (1) CL trains self-attentions to capture longer-range global patterns than MIM, such as the shape of an object, in the later layers of the ViT architecture. This CL property helps ViTs linearly separate images in their representation spaces. However, it also makes the self-attentions collapse into homogeneity for all query tokens and heads. Such homogeneity of self-attention reduces the diversity of representations, worsening scalability and dense prediction performance. (2) CL utilizes the low-frequency signals of the representations, but MIM utilizes high-frequencies. Since low- and high-frequency information respectively represent shapes and textures, CL is more shape-oriented and MIM more texture-oriented. (3) CL plays a crucial role in the later layers, while MIM mainly focuses on the early layers. Upon these analyses, we find that CL and MIM can complement each other and observe that even the simplest harmonization can help leverage the advantages of both methods.

1 INTRODUCTION

Contrastive Learning (CL) (He et al., 2020; Chen et al., 2020a;b; 2021) has been the most popular self-supervised learning methods until recently. It aims to learn the invariant semantics of two random views (Tian et al., 2020a;b) by making global projections of representations similar for positive samples and dissimilar for negative samples. Since CL exploits the globally projected representations to contrast each other, it can be deemed as an *"image-level"* self-supervised learning approach.

Deviating from CL, masked image modeling (MIM) (Bao et al., 2022; Xie et al., 2022; He et al., 2022) has risen as a strong competitor of CL in the era of Vision Transformers (ViTs) (Dosovitskiy et al., 2021) with its impressive performances of downstream tasks. MIM trains ViTs by reconstructing the correct semantics of masked input patches. Unlike CL, it learns the semantics of patch tokens and this can be deemed as a *"token-level"* self-supervised learning approach. Since MIM outperforms CL in fine-tuning accuracy, it may appear *prima facie* as a more effective pre-training method than CL. However, a different trend is observed for linear probing accuracy with CL outperforming MIM (see Figure 1). For further exposition on CL and MIM, we refer the reader to Appendix B.

Then, which method—CL or MIM—should we use for the self-supervised learning of ViTs? Although both methods are widely used, little is known about what they learn. This paper sheds light on their nature by showing that ViTs trained through CL and MIM learn opposite knowledge. In particular, we raise questions to better understand self-supervised learning, and then find the answers that can potentially affect future improvements. The questions posed can be divided into the following properties of Vision Transformers: the behavior of self-attentions, the transformation of the representations, and the position of lead role components. Our key questions and findings are elaborated below.

How do self-attentions behave? (Section 2) We find that CL primarily captures global relationships, while MIM captures local relationships. This implies that the representations of CL contain more global patterns, such as object shapes, than those of MIM. On the one hand, this property helps CL recognize objects and distinguish images. On the other hand, however, it also suggests that CL struggles to preserve local information that ViTs trained with CL lack such information. In particular, we observe that self-attentions of CL in the later layers for all query tokens and heads collapse into

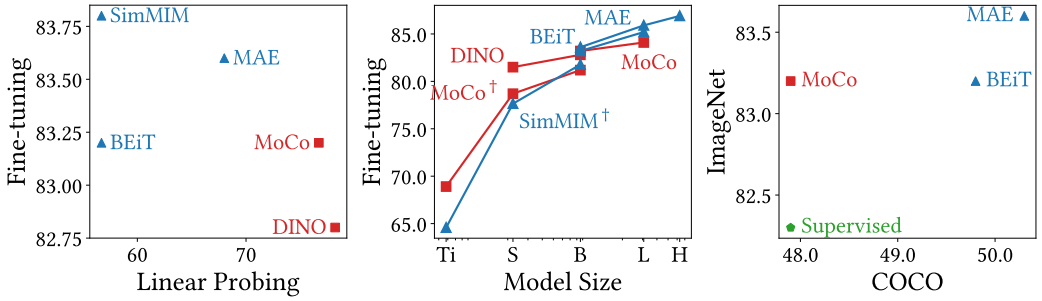


Figure 1: **CL outperforms MIM in linear probing and small model regimes.** In contrast, MIM excels in fine-tuning, large model regimes, and dense prediction. Red squares (■) denote CL, and blue triangles (▲) denote MIM. By default, we report the performance of ViT-B trained or pre-trained on ImageNet-1K. We use the results from original papers and (He et al., 2022) for object detection. Regarding the scaling experiment, we report the results that we reproduced based on official configurations except with 100 epochs, marking them as MoCo† and SimMIM†. *Left*: CL outperforms MIM in linear probing but underperforms in fine-tuning. *Middle*: CL outperforms MIM in small model regimes (ViT-Ti and ViT-S), and MIM shows superior scalability in large model regimes (ViT-L and ViT-H). *Right*: MIM outperforms CL in the dense prediction downstream tasks, such as object detection with Mask R-CNN (He et al., 2017) on COCO (Lin et al., 2014).

homogeneous attention maps. In such cases, most self-attention maps focus on object boundaries, meaning that they can capture object shapes but may lose interaction diversity of self-attention. Consequently, CL and MIM each have advantages over different tasks: CL works well for linear probing and classification tasks with smaller models, whereas MIM outperforms CL in fine-tuning and dense prediction tasks with larger models.

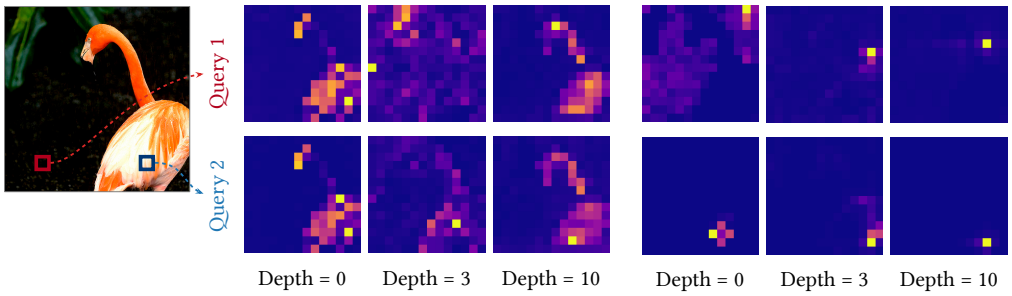
How are representations transformed? (Section 3) CL transforms representations mainly based on image-level information, and its self-attentions collect information on object shape over entire tokens. This process makes tokens similar rather than diversifying them. As a result, CL distinguishes images well but has difficulty distinguishing tokens. On the contrary, MIM preserves and amplifies token-level information. Thus, the self-attentions for each token are substantially different and prohibit each token from including redundant information. We observe the consistent property from our Fourier analysis: CL primarily utilizes the low-frequency signals, but MIM utilizes high-frequencies. This observation suggests that CL is shape-biased and MIM is texture-biased. In sum, self-supervised models trained with CL and MIM learn the representations in different levels of detail.

Which components play an important role? (Section 4) Analyses of the importance of each CL and MIM layer demonstrate that the later layers in CL and early layers in MIM play a key role. We interpret this as a consistent observation since early layers are usually known to capture low-level features—e.g., local patterns, high-frequency signals, and texture information—and later layers capture global patterns, low-frequency signals, and shape information (Dosovitskiy et al., 2021; Raghu et al., 2021; d'Ascoli et al., 2021; Graham et al., 2021; Dai et al., 2021; Park & Kim, 2022b).

From the above analyses and insights, we find that CL and MIM can complement each other and show in Section 5 that even the simplest implementation, such as a linear combination of CL and MIM objectives, can take advantage of both methods. Surprisingly, the hybrid models outperform those pre-trained with either CL or MIM both in terms of fine-tuning and linear probing accuracy.

2 How Do Self-Attentions Behave?

We point out that CL and MIM may not be silver bullets for all tasks by visualizing their performance from three perspectives. First, Figure 1 shows that CL generally outperforms MIM in linear probing, while MIM dominates CL in the fine-tuning scheme. However, when we dissect the size of the model, CL outperforms MIM after fine-tuning for small models (cf. (Wang et al., 2022)), while MIM performs better on large models. Also, MIM yields effective representations for dense prediction tasks, such as object detection, but CL falls short on those tasks. This section explains these phenomena by investigating the behavior of self-attentions.



(a) MoCo (b) SimMIM

Figure 2: **Self-attentions of CL (MoCo) capture global relationships, but they collapse into homogeneous attention maps for all query tokens and heads.** Self-attentions of MIM (SimMIM) focus on local areas. We visualize the attention maps for two different query tokens in the beginning through the end layers. We omit the results for self-attention heads, which show mostly consistent results. *Left*: Self-attentions of CL capture global patterns and the shape of an object. However, all attention maps capture the same shape information regardless of the query tokens. *Right*: Self-attentions of MIM capture local patterns, and they correlates with the query tokens.

Our analyses mainly compare ViT-B/16 pre-trained on ImageNet-1K (Russakovsky et al., 2015) with MoCo v3 (Chen et al., 2021) and SimMIM (Xie et al., 2022). We use the ImageNet validation images for the experiments. We observe that other methods, e.g., DINO (Caron et al., 2021), BEiT (Bao et al., 2022), and MAE (He et al., 2022), have consistent properties (See Figure B.1).

CL mainly captures global relations, but MIM does not. We measure the ranges of self-attentions via attention distance (Dosovitskiy et al., 2021). Attention distance is defined as the average distance between the query tokens and key tokens considering their self-attention weights, and it conceptually corresponds to the receptive fields in CNNs.

Figure 3 shows that the attention distance of CL (MoCo) is significantly higher than that of MIM (SimMIM), especially in the later layers. Considering the qualitative visualization, Figure 2, this implies that the representations of CL contain global patterns and shape information, so it can help ViT distinguish between objects or images. Conversely, the self-attentions of MIM mainly capture local relationships, i.e., MIM may have difficulty recognizing whole objects and their shapes. Section 3 also discuss this claim from a representational perspective.

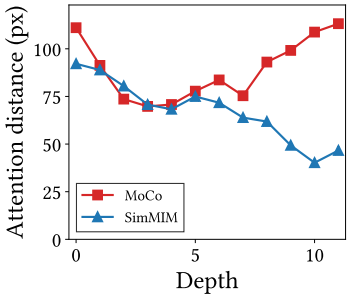


Figure 3: **The effective receptive fields of CL are global, but those of MIM are local.** Such phenomenon is conspicuously observed in later layers.

Self-attentions of CL collapse into homogeneity. We observe an interesting behavior of CL in Figure 2, which shows attention maps for query tokens from two different spatial locations. Surprisingly, the later self-attentions of CL indicate *almost identical* object shapes for the two query tokens, compared to that of MIM. We describe this phenomenon as *attention collapse into homogeneity*. This attention collapse trend in self-attentions of CL is observed across all the heads and query tokens. In contrast, MIM shows more faithful self-attentions for two query tokens, as expected.

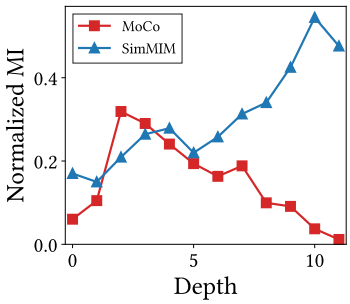
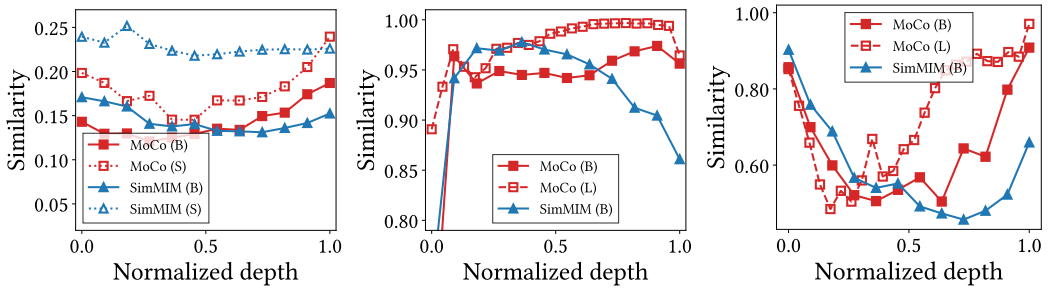


Figure 4: **Self-attentions of CL have little to do with query tokens.** CL's normalized mutual information is significantly lower than that of MIM in the later layers.

We use normalized mutual information (NMI) (Strehl & Ghosh, 2002) to measure the attention collapse. Let $p(q)$ be a distribution of query tokens, and assume they are uniformly distributed since a single query token is given for each spatial coordinate, i.e., $p(q) = {}^{1}/_{N}$ where N is the number of the tokens. Then the joint distribution of query and key tokens is $p(q, k) = \pi(k|q)p(q)$ where $\pi(k|q)$ is softmax-



(a) Between heads (b) Between depths (c) Between tokens

Figure 5: **CL lacks representational diversity in the later layers.** We measure cosine similarities of representations in self-attentions between the heads (left), depths (middle), and spatial coordinates (right). All the results show that the representational similarity of later self-attentions of CL is higher than that of MIM. Increasing the model size of CL is not effective in improving the diversity. *Left*: The similarity of representations from two heads in self-attention. *Middle*: The similarity between representations before and after self-attentions transform them. *Right*: The similarities of representations at two spatial coordinates. ViT-L is trained with 25 epochs.

normalized self-attention matrix. Thus, the normalized mutual information is $\frac{I(q,k)}{\sqrt{H(q)H(k)}}$ where $I(\cdot,\cdot)$ is the mutual information and $H(\cdot)$ is the marginal entropy. The low mutual information value represents the attention map is less dependent on the query tokens implying the attention collapse into homogeneity. Conversely, the high the mutual information means the attention maps strongly depend on the query token.

Figure 4 shows the degree of attention collapse in terms of the normalized mutual information (NMI). The results show that the mutual information of CL is significantly lower than that of MIM in the later layers, suggesting that self-attentions of CL tend to collapse into homogeneous distributions.

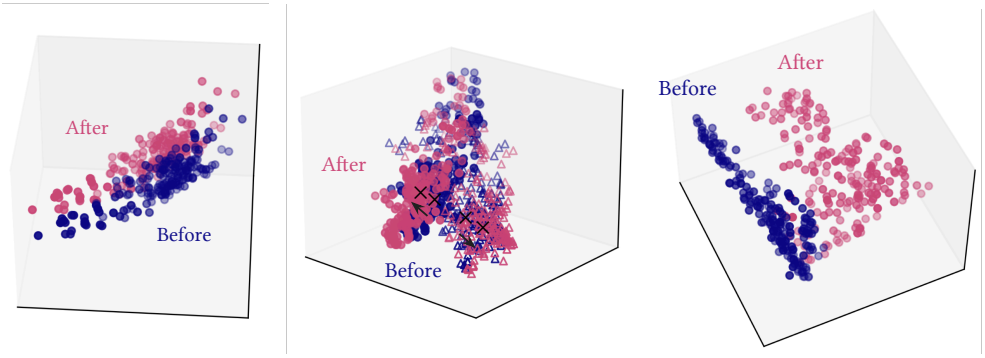
The attention collapse reduces representation diversity. We conjecture that the self-attention collapse into homogeneity eventually leads to homogeneous token representations. To support this argument, we measure representational cosine similarities. In particular, we design three similarities: between different self-attention heads (heads), between the before and after self-attention layers (depths), and between different tokens (tokens).

Figure 5 shows the result reporting the representation cosine similarities for heads, depths, and tokens. As expected, the similarities of CL are notably higher than those of MIM in the later layers, indicating that the representations of CL have significant homogeneity. Moreover, increasing the model size does not solve or even exacerbates this problem of CL. Increasing the number of heads (ViT-S to ViT-B; Figure 5a) improves the diversity of MIM due to its self-attention diversity, but hardly improves CL. Increasing the depth of CL (ViT-B to ViT-L; Figures 5b and 5c) only adds redundant modules.

Implications of the behaviors we observed. In conclusion, the self-attention of CL captures global patterns and shapes of objects. However, CL suffers from the problem of attention collapse into homogeneity, which reduces the diversity of token representations. On the contrary, MIM does not suffer from the attention collapse problem by primarily capturing local patterns.

The behaviors mentioned above can explain the phenomena we observed in Figure 1:

- CL outperforms MIM in linear probing tasks because capturing shapes by CL helps recognize objects and distinguish images. Although MIM preserves texture and the diversity of representation, these may not strongly correlate with objects or content, compared to shapes.
- The attention collapse prohibits CL from fully exploiting heads, depths, and tokens of ViTs. Since homogeneous representations is not very helpful for improving token representations, ViTs trained with CL wastes a large part of network capability. Therefore, the fine-tuning accuracy of MIM with large models is significantly higher than that of CL.
- CL is not suitable for dense prediction since the token features are homogeneous with respect to their spatial coordinates.



(a) MoCo (one image) (b) MoCo (two images) (c) SimMIM (one image)

Figure 6: **Self-attention layers of CL and MIM transform representations differently.** We visualize 196 spatial representational tokens for an example validation image in representation space. The blue (•) and red (•) data points stand for the tokens before and after the first self-attention transformation. *Left*: The self-attentions of CL (e.g., MoCo) translate all the tokens equally, so the distances between tokens in an image do not change. *Middle*: However, CL makes the "centers of representations (represented by ×)" away from each other. Therefore, the image is linearly separable. The circle (•) and triangle (△) data represent tokens from different images. *Right*: The self-attentions of MIM (e.g., SimMIM) differently transform representations according to query tokens. Thus, the distances between tokens change. See Figure 7 for quantitative analyses.

We further investigate the self-attention's behavior with restricted receptive fields in Figure A.1. As shown in the experiment, locally restricted self-attentions lead to lower linear probing but higher fine-tuning accuracy, which is consistent with our observations.

3 How Are Representations Transformed?

In this section, we analyze the token representations of ViTs pre-trained with CL and MIM to demonstrate how the properties of self-attentions we observed in Section 2 affect the representations differently. We use the same pre-trained ViT-B/16 models by default as well as we did in Section 2.

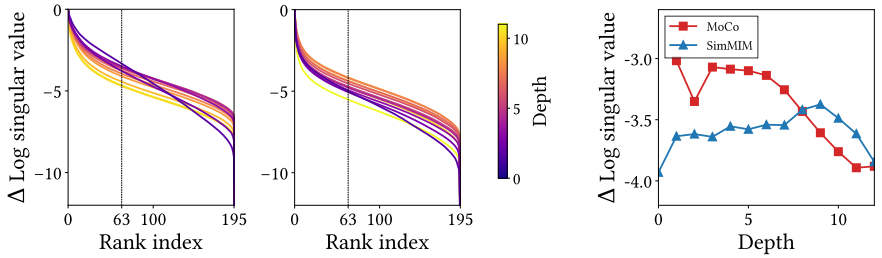
CL transforms all tokens in unison, while MIM's tokens are transformed individually. To show how CL and MIM transform token representations, we visualize them in representation space. Figure 6 represents 196 (14×14 patches) tokens before and after the first self-attention module from a single image of the ImageNet validation set. We use the top three singular vectors obtained via singular value decomposition (SVD) as the bases of the representation space.

Figure 6a visualizes the changes in the tokens of CL transformed by the self-attention module; it indicates that the self-attentions of CL translate all tokens in unison. This phenomenon occurs since the self-attention maps of CL are homogeneous, i.e., the self-attention is almost independent of the spatial coordinates and the query tokens. Therefore, the module adds near-constant to all the token representations. As a result, the inter-representation distance and the representations' volume remain almost invariant, which implies that CL cares less about individual tokens.

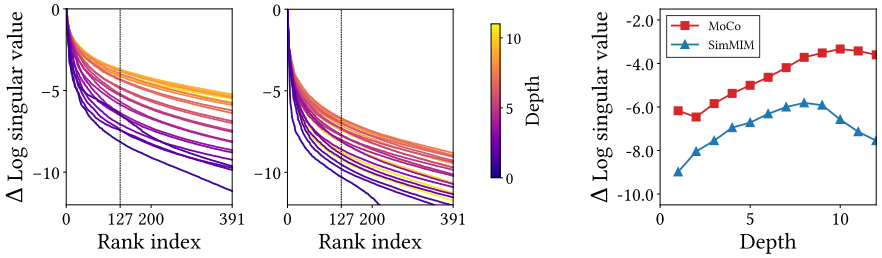
Nevertheless, self-attentions are essential for the discriminative power of CL. As shown in Figure 6b, they distinguish images by moving "the centers of the representation distribution" away from each other. In short, this figure suggests that CL makes the image linearly separable even though it loses the ability to distinguish tokens.

In contrast, MIM applies a different transform to individual tokens, as shown in Figure 6c, because the different self-attentions are assigned to the individual spatial tokens. Thus, MIM alters the distance between the tokens of a single image and the volume of the representation distribution.

We find consistent results in quantitative analysis. Inspired by Jing et al. (2022), Figure 7 visualize singular value spectra for tokens and images. The singular value spectrum provides singular values of representation distribution obtained by SVD, so we can use it to represent the effective volume



(a) Singluar value spectum of tokens from a single image



(b) Singluar value spectum of images

Figure 7: **CL does not change or even decreases the distribution volume of tokens from a single image,** implying that it hardly distinguishes each token. Instead, it significantly increases the distribution volume of images. To demonstrate these properties, we visualize singular value spectra, the singular values of the distribution of representation sorted by the magnitude. The higher the singular value, the larger the volume of the distribution. The right of this figure is the 64th and 128th highest singular value for depth. *Top*: Singular value spectrum of tokens from a single image. CL decreases the singular values of the tokens, but MIM increases. *Bottom*: Singular value spectrum of images. CL significantly increases the volumes occupied by images, but MIM hardly increase the volume.

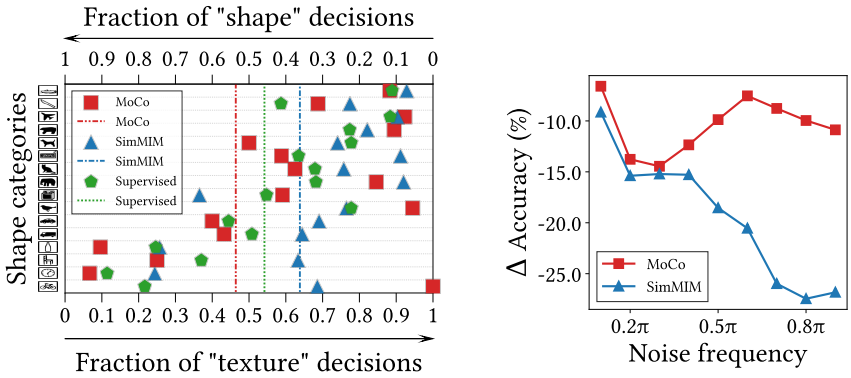
of distribution in representation space. The higher the singular value in the spectrum, the larger the volume of the representation distribution. To calibrate the scale of volumes, we use the relative log singular value (Δ Log singular value), the difference with the largest singular value for a depth.

Figure 7a shows the singular value spectra of tokens for a single image. We calculate it for each image in ImageNet validation set and report averaged singular values over the dataset. In this figure, the layers of CL hardly increase or even decrease the singular value; consistent with the explanation above, it implies that CL hardly distinguishes tokens. In contrast, MIM increases the singular value, so it changes the volume of tokens and can distinguish tokens. One interesting observation is that a few later layers of MIM decrease the volume, even though they captures local patterns as shown in Figures 3 and 4. This is because they behave like decoders, and Section 4 discuss this in detail.

Figure 7b shows the singular value spectra of images. We averaged all tokens in an image to build an image-level representation vector and conduct the singular value spectra over the collection of representations in the validation set. As opposed to the previous case, the representational volume of CL is larger than that of MIM, which implies that CL makes the image-level representation separable.

CL exploits low-frequencies, and MIM exploits high-frequencies. We hypothesize that CL captures low-frequency and MIM captures high-frequency information in spatial dimensions since CL provides image-level self-supervision to capture global patterns, while MIM provides token-level self-supervision to exploit local patterns. To support this argument from a frequency perspective, we conduct Fourier analysis on the representations as following Park & Kim (2022b). In particular, we report the relative log amplitude of Fourier transformed representation by calculating the amplitude difference between the representations' highest and lowest frequencies.

Figure 9 visualizes the relative amplitudes of CL and MIM. It shows that the high-frequency amplitude of CL is significantly smaller than that of MIM, suggesting that CL mainly utilizes low-frequency spatial information such as global structures and shapes. On the contrary, MIM usually uses high-frequency spatial information such as narrow structures and fine textures.



(a) Stylized ImageNet (b) Robustness for noise frequency

Figure 8: **CL is biased toward shape, whereas MIM is biased toward texture.** We report the predictive results of models for linear probing tasks. However, we observe consistent results for the fine-tuned models (See Figure C.1b). *Left:* The classification results on Stylized ImageNet. It shows that CL is more shape-biased than MIM and even than the supervised pre-trained model. The vertical lines represent the averaged results for the shape categories. We also report the results of supervised ViT with ImageNet-1K class labels for comparison. *Right:* The accuracy drop on images with frequency-based random noises. MIM shows a more significant amount of accuracy drop with high-frequency noises than CL, demonstrating MIM's texture-biased property compared to CL. The frequency window size is 0.1π for the frequency-based noise.

Another interesting finding is that the last few layers of MIM reduce the high frequencies even though they only focus on the local areas (See Figure 3). We conjecture that MIM implicitly divides ViTs into the encoder-decoder structure and lets the intermediate layers have linearly separable information. In contrast, CL lets the last layer have it. This is further elaborated in Figure 11.

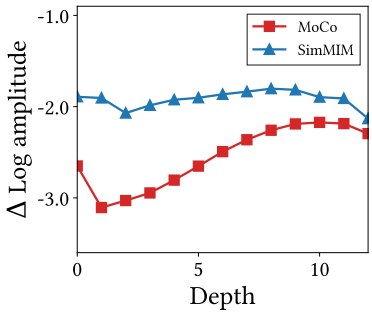


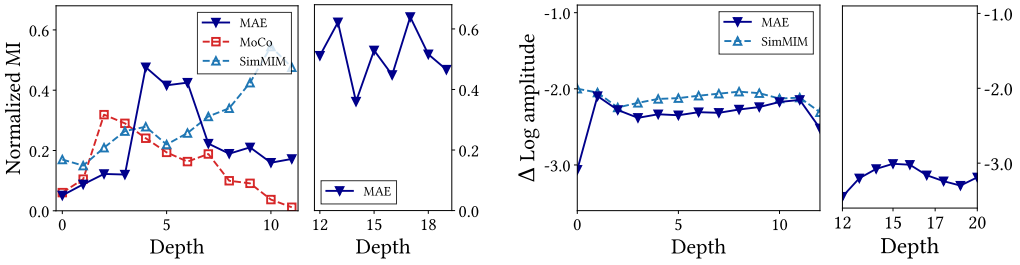
Figure 9: **CL exploits low-frequency, but MIM exploits high-frequency**. Moreover, a few last layers of CL reduce high-frequency by capturing global patterns. MIM also reduces it even though they capture local patterns, because the later layers behave like decoders. See also Figure 11.

CL is shape-biased, but MIM is texture-biased. Based on the results of Fourier analysis, we assume that CL and MIM have a bias toward shape and texture, respectively. To demonstrate this claim, we use Stylized-ImageNet (Geirhos et al., 2019), a texture-altered dataset, by using AdaIN (Huang & Belongie, 2017). Figure 8a report the linear probing results on Stylized-ImageNet to evaluate the shape and texture biases of pre-trained models. Compared to the model pre-trained with supervised learning, CL depends more on the shape of images to classify images, and MIM depends on texture. In other words, CL is robust to texture changes, and MIM is vulnerable to them.

Figure 8b shows the consistent results. In this experiment, we follow Park & Kim (2022a;b) and measure the decrease in accuracy on the ImageNet dataset with frequency-based random noise. The results suggest that CL is robust to high-frequency noises, but MIM is significantly more vulnerable to them. Since high-frequency noises harm the fine details of images, we arrive at the same conclusion that CL is more shape-biased and MIM is texture-biased. This could serve as an explanation for the robustness of CL against adversarial perturbations (Bordes et al., 2021).

4 Which Components Play an Important Role?

In the previous sections, various perspectives consistently show that CL exploits image-level global patterns and MIM captures token-level local patterns. This section analyzes pre-trained ViTs from the architecture perspective layer-wisely and shows key components in CL and MIM are different.



(a) Self-attention (b) Fourier analysis

Figure 10: **The explicit decoder architecture of MAE helps ViT effectively leverage the advantages of MIM.** We analyze the encoder and the decoder in terms of self-attention and Fourier analysis by using the model for pre-training with masking ratio of zero. The left of each figure represents the encoder and the right represents the decoder. *Left:* The mutual information of MAE is lower than that of SimMIM in the encoder but higher in the decoder. *Right:* Later layers of MAE except the last layer do not reduce high-frequency information whereas that of SimMIM reduce it. In particular, the decoder captures low-frequency and the encoder captures high-frequency compared with each other.

Later layers of CL and early layers of MIM are important. According to studies on ViT architectures (Graham et al., 2021; Dai et al., 2021; Park & Kim, 2022b), the later layers use high-level information, and the early layers exploit low-level information. Since CL and MIM exploit global and local patterns, we expect that the later layers of CL and early layers of MIM play key roles.

To evaluate the importance of each layer, we measure the linear probing accuracy using the intermediate representations with the configuration of Table A.1. In Figure 11, we observe the following properties: First, the linear probing accuracy of MIM is higher than that of CL at the beginning. Conversely, CL outperforms MIM at the end of the model. This result indicates that later layers of CL and early layers of MIM play an important role in making linearly separable representations. Second, the accuracy of CL increases with increasing depth as expected, but surprisingly, the accuracy of MIM decreases at the end of the model, i.e., later layers of MIM are not very helpful in separating representations. We explain this as MIM methods with a shallow prediction head, e.g., SimMIM, use later layers of the backbone as a decoder. Therefore, MIM with a deep self-attention decoder, e.g., MAE (He et al., 2022), can be useful for linear probing performance. Moreover, it also explains SimMIM's high-frequency component and representational volumes drop at later layers as shown in Figures 7 and 9. Third, the highest linear probing accuracy of MIM is lower than that of CL.

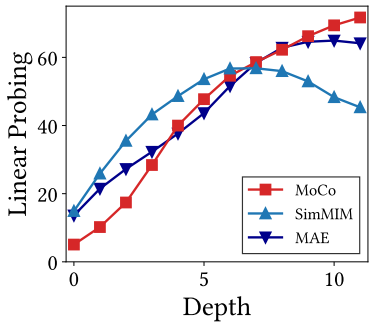
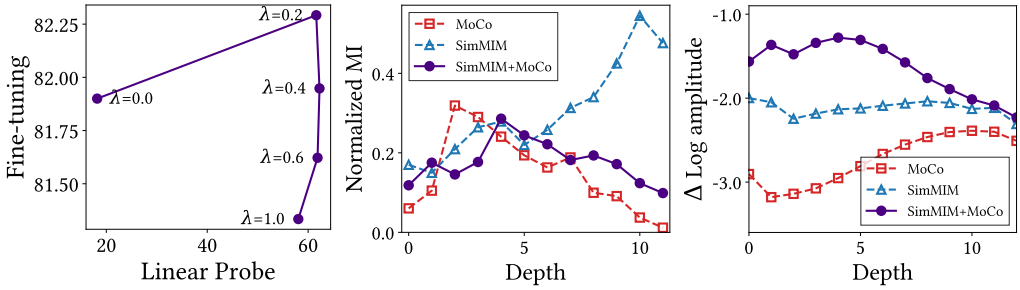


Figure 11: **Later layers of CL and early layers of MIM play a key role.** We report linear probing accuracies by using representations of the intermediate layer. CL outperforms MIM in later layers, and MIM outperforms CL in early layers.

The explicit decoder help further leverage the advantages of MIM. Several previous observations find that implicit decoders in MIMs with shallow prediction heads, such as SimMIM, can impair performance. MAE (He et al., 2022) addresses this problem by introducing deep explicit ViT decoders and using masked tokens only for the decoders.

In Figure 10, we analyze the token representations of MAE to understand the properties of decoders more deeply. Figure 10a represents the self-attention behavior. This result indicates that the mutual information of MAE is lower than that of SimMIM in the later encoder layers but higher in the decoder, implying that decoders and self-attentions are correlated.

Figure 10b shows the result of Fourier analysis. The later layers (except the last layer) of MAE do not reduce high frequency component, whereas the last four layers of SimMIM reduce it. The decoder of MAE utilizes low-frequency information compared with the encoder. As a result, the backbone of MAE does not have an unnecessary decoder part to learn effective token representations, and it makes efficient use of the limited capacity (cf. (Wang et al., 2022)).



(a) Performance (b) Self-attention (c) Fourier analysis

Figure 12: **The simple linear combination of CL (MoCo) and MIM (SimMIM) objectives outperforms the vanilla CL and MIM**. λ is importance weight of CL, so $\lambda = 0$ means SimMIM and $\lambda = 1$ means MoCo. *Left:* "CL + MIM" outperforms CL and MIM in both linear probing and fine-tuning accuracy. *Middle:* The mutual information of "CL + MIM" decreases at the end of the model, suggesting that the self-attentions of later layers collapse into homogeneity and capture the same object shape information. *Right:* Fourier analysis shows that "CL + MIM" amplifies high-frequencies at the beginning and reduces them at the end. It implies that "CL + MIM" exploits high-frequency information at the beginning and low-frequency information at the end.

5 Can We Benefit from Both Methods?

We present comparative analyses on CL and MIM from three perspectives: self-attentions, representation transforms, and the position of important layers. All of our results indicate that CL and MIM train ViTs differently. These differences naturally imply that combining CL and MIM may help leverage the advantages of both methods.

To show that CL and MIM are complementary, we introduce the simplest way to harmonize CL and MIM by linearly combining two losses, i.e., $\mathcal{L} = (1-\lambda)\mathcal{L}_{\text{MIM}} + \lambda\mathcal{L}_{\text{CL}}$ where $\mathcal{L}_{\text{MIM}}$ and $\mathcal{L}_{\text{CL}}$ each indicate the losses of MIM and CL, and λ is the importance weight of CL. We find that this simple hybrid model trained with combined losses efficiently exploits the strengths of both methods. Figure 12a shows linear probing and fine-tuning accuracy on ImageNet with varying λ. Surprisingly, the hybrid models outperform MIM ($\lambda = 0$) and CL ($\lambda = 1$) in both aspects. Figure 12b and Figure 12c can provide insights on how hybrid models behave by analyzing the model with $\lambda = 0.2$ in terms of self-attentions in Section 2 and Fourier analysis in Section 3, respectively; both results show that the hybrid model exploits MIM properties in the early layers and CL's properties in the later layers. In particular, Figure 12b indicates that the self-attentions of the early layers are changed according to the query token but those of the later layers are not. Likewise, Figure 12c shows that the early layers exploit high-frequency, while the later layers try to exploit low-frequency.

6 Conclusion

We have conducted comparative studies highlighting various facets of two widely used self-supervised learning methods for vision transformers: contrastive learning and masked image modeling. The studies demonstrated many opposing properties of two methods: image information (image-level *vs*. token-level; as in Section 2), feature representations (low-frequency *vs*. high-frequency; as in Section 3), and different lead role components (later layers *vs*. early layers; as in Section 4). We further imagined the possible application exploiting only the benefits from both ends and showed how their mixture outperforms them.

Future directions. Based on our study, diverse future directions can be explored. We believe there are better ways beyond a simple linear combination of CL and MIM objectives. For example, novel self-supervised learning can be considered, which applies CL for the later layers and MIM for the early layers. Moreover, we may extend our findings on self-supervision for multi-stage ViTs, such as PiT (Heo et al., 2021) and Swin (Liu et al., 2021). Another interesting direction is to enhance the properties of CL or MIM. Techniques that help CL or MIM learn shapes or textures also may improve performance.

REPRODUCIBILITY STATEMENT

We will provide the code and experimental details for reproducing the results. The code will be released as open source. See Appendix A for detailed configurations.

REFERENCES

Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. In *International Conference on Learning Representations*, 2022.

Florian Bordes, Randall Balestriero, and Pascal Vincent. High fidelity visualization of what your self-supervised representation knows about. *arXiv preprint arXiv:2112.09164*, 2021.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *International Conference on Computer Vision*, 2021.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, 2020a.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021.

Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020b.

Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *International Conference on Computer Vision*, 2021.

Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Advances in Neural Information Processing Systems*, 2020.

Zihang Dai, Hanxiao Liu, Quoc V Le, and Mingxing Tan. Coatnet: Marrying convolution and attention for all data sizes. *Advances in Neural Information Processing Systems*, 2021.

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. 2021.

Stéphane d'Ascoli, Hugo Touvron, Matthew L Leavitt, Ari S Morcos, Giulio Biroli, and Levent Sagun. Convit: Improving vision transformers with soft convolutional inductive biases. In *International Conference on Machine Learning*, pp. 2286–2296. PMLR, 2021.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*, 2019.

Benjamin Graham, Alaaeldin El-Nouby, Hugo Touvron, Pierre Stock, Armand Joulin, Hervé Jégou, and Matthijs Douze. Levit: a vision transformer in convnet's clothing for faster inference. In *International Conference on Computer Vision*, 2021.

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. *Advances in neural information processing systems*, 2020.

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

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022.

Byeongho Heo, Sangdoo Yun, Dongyoon Han, Sanghyuk Chun, Junsuk Choe, and Seong Joon Oh. Rethinking spatial dimensions of vision transformers. In *International Conference on Computer Vision*, 2021.

Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q Weinberger. Deep networks with stochastic depth. In *European Conference on Computer Vision*, 2016.

Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, 2017.

Li Jing, Pascal Vincent, Yann LeCun, and Yuandong Tian. Understanding dimensional collapse in contrastive self-supervised learning. In *International Conference on Learning Representations*, 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 *European conference on computer vision*, 2014.

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 *International Conference on Computer Vision*, 2021.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Namuk Park and Songkuk Kim. Blurs behave like ensembles: Spatial smoothings to improve accuracy, uncertainty, and robustness. In *International Conference on Machine Learning*, 2022a.

Namuk Park and Songkuk Kim. How do vision transformers work? In *International Conference on Learning Representations*, 2022b.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, 2019.

Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? *Advances in Neural Information Processing Systems*, 2021.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 2015.

Alexander Strehl and Joydeep Ghosh. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *Journal of machine learning research*, 2002.

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*, 2016.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *European Conference on Computer Vision*, 2020a.

Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? *Advances in Neural Information Processing Systems*, 33:6827–6839, 2020b.

Yuandong Tian, Xinlei Chen, and Surya Ganguli. Understanding self-supervised learning dynamics without contrastive pairs. In *International Conference on Machine Learning*, 2021.

Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8769–8778, 2018.

Shaoru Wang, Jin Gao, Zeming Li, Jian Sun, and Weiming Hu. A closer look at self-supervised lightweight vision transformers. *arXiv preprint arXiv:2205.14443*, 2022.

Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Dense contrastive learning for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022a.

Yixuan Wei, Han Hu, Zhenda Xie, Zheng Zhang, Yue Cao, Jianmin Bao, Dong Chen, and Baining Guo. Contrastive learning rivals masked image modeling in fine-tuning via feature distillation. *arXiv preprint arXiv:2205.14141*, 2022b.

Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022.

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 *International Conference on Computer Vision*, 2019.

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

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(3):302–321, 2019.

Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. *International Conference on Learning Representations*, 2022.

A SETUP

We build the configurations based on Xie et al. (2022) for fine-tuning and pre-training, and Caron et al. (2021) for linear probing. For data augmentation and regularization, we adopt widely used settings, e.g., Randaugment (Cubuk et al., 2020), label smoothing (Szegedy et al., 2016), mixup (Zhang et al., 2018), cutmix (Yun et al., 2019), stochastic depth (Huang et al., 2016). Layer decay (Bao et al., 2022) is also used. Table A.1 summarize the configurations. We use {1, 4, 8} NVIDIA A100 Tensor Core GPU for all experiments. Neural network models are implemented in PyTorch(Paszke et al., 2019); The code will be released publicly.

Table A.1: **Training settings.**

CONFIGURATION	Linear Probing	Fine-tuning	Pre-training
`optimizer`	sgd	adamw	adamw
`base learning rate`	1.0e-0	1.25e-3	1e-4
`weight decay`	0.05	0.05	0.05
`batch size`	1k	2k	1k
`training epoch`	50	100	100
`learning rate schedule`	cosine	cosine	multistep
`warmup epoch`	0	20	10
`warmup schedule`	·	linear	linear
`randaugment`	·	9, 0.5	9, 0.5
`label smoothing`	·	0.1	0.1
`mixup`	·	0.8	0.8
`cutmix`	·	1.0	1.0
`stochastic depth`	·	0.1	0.1
`layer decay`	·	0.65	1.0
`gradient clip`	·	5.0	5.0

B RELATED WORK

CL is a method based on comparing the global projection of two different random views. However, this approach usually suffers from the collapsing problem, where all representations collapse into constant solutions. To solve this problem, various methods such as negative samples and InfoNCE (Oord et al., 2018) have been proposed. Negative samples is an effective technique to avoid the collapsing problems, but they cause dimensional collapse (Jing et al., 2022) and require extra large batches (Chen et al., 2020a) or memory queues (He et al., 2020; Chen et al., 2020b) to retrieve them.

Some SSL methods, e.g. Grill et al. (2020); Caron et al. (2021), do not use negative samples and use the projections of their Siamese representations as the positives. Such self-distillation has been explored theoretically and empirically (Chen & He, 2021; Tian et al., 2021) to prevent the collapsing problem, but we do not discuss the distillation scheme in this paper. Wei et al. (2022b) shows that feature distillation improves the fine-tuning performance of CL by diversifying attention ranges; this observation is consistent with our

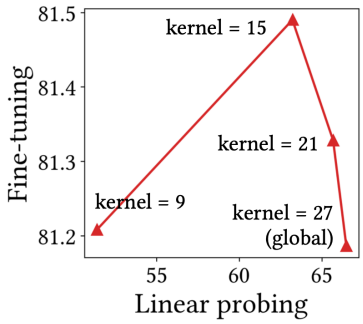
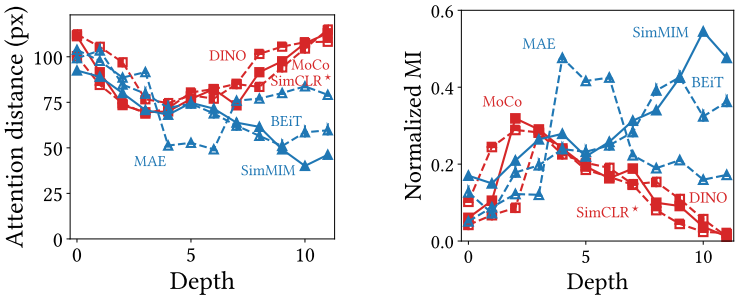
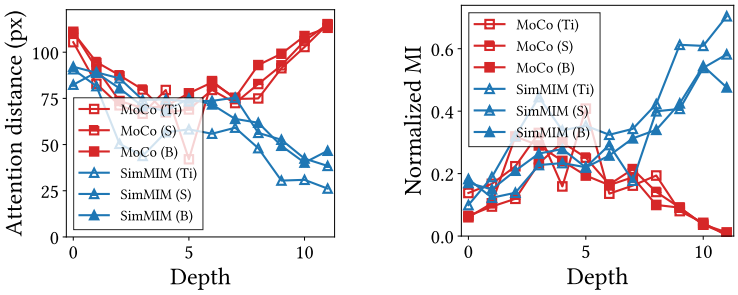


Figure A.1: **Restricted self-attention harms linear probing accuracy but improves fine-tuning accuracy.** We report the fine-tuning accuracy and linear probing accuracy of ViTs with restricted self-attention via attention mask.



(a) Attention distance (b) Normalized MI

Figure B.1: **MIM methods and CL methods have consistent properties.** We visualize the behaviors of self-attention in terms of attention distance and normalized mutual information (MI). SimCLR* stands for MoCo with the momentum coefficient of 0, introduced in Chen et al. (2021). *Left:* The attention distance of CL methods, namely MoCo, SimCLR* and DINO, is higher than that of MIM methods, namely SimMIM, BEiT, and MAE. It suggests that CL methods consistently capture global patterns. *Right:* The normalized mutual information of MIM is higher than that of CL, i.e., the self-attentions of MIM are correlated with query tokens.



(a) Attention distance (b) Normalized MI

Figure B.2: **ViTs have consistent self-attention behaviors regardless of size.** We visualize the self-attention behaviors of three ViTs—ViT-{Ti, S, B}—in terms of attention distance and normalized mutual information (MI). We observe consistent properties from small to large ViTs. *Left:* All self-attentions of MoCo capture global patterns in the later layers. In contrast, SimMIMs capture global pattern. *Right:* Likewise, all later self-attention maps of MoCo collapse into homogeneity.

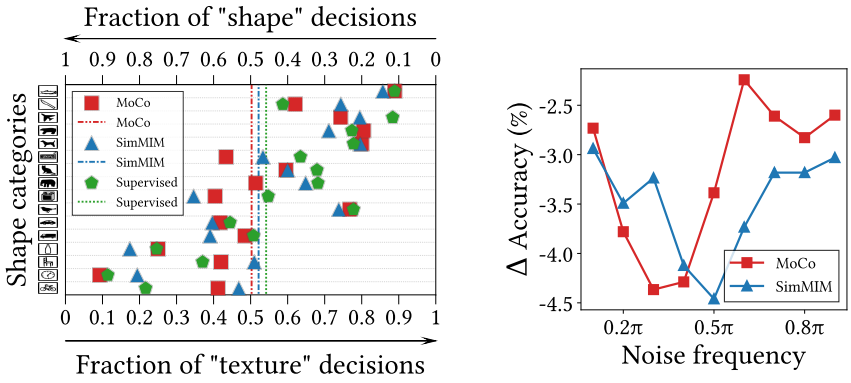
findings. While they focus on distillation to improve CL, we reveal the fundamental nature of self-supervised learning by rigorously comparing CL and MIM.

Compared with CL, MIM has been rarely explored. Various methods, such as histograms of oriented gradients (Wei et al., 2022a) and tokenization (Bao et al., 2022), have been proposed as part of porting masked language models to the image domain with ViTs. Among them, SimMIM (Xie et al., 2022) and MAE (He et al., 2022) are simple yet effective methods to reconstruct masked tokens without complicated pretext tasks. Because of its simplicity and superior performance in downstream operations, MIM is attracting attention as a promising technique in image processing.

Nevertheless, we find hints suggesting that CL and MIM utilize different aspects of the data, making them complementary. For example, Zhou et al. (2022); Wang et al. (2021); Yu et al. (2022) achieve high predictive performance by harmonizing the image-level and the token-level self-supervised learning.

C Restricted Self-Attention

In Section 2, we demonstrate that the homogeneity of self-attention map, due to the attention collapse, helps ViT distinguish images but harms fine-tuning accuracy. One simple method to inject locality into self-attentions is to limit the receptive field of self-attention by using attention masks. Figure A.1



(a) Stylized ImageNet (b) Robustness for noise frequency

Figure C.1: **Fine-tuned ViT inherits the robustness property against frequency-based noises.** Similar to Figure 8b, we measure the decrease in accuracy of ViTs fine-tuned with CL and MIM. *Left:* Even with fine-tuned ViTs, CL is shape-biased and MIM texture-biased when compared to each other. However, the performance gap is smaller than the result for the linear probing models. *Right:* Robustness against frequency-based random noise suggests the same conclusion: CL is robust against high-frequency noises, but MIM is not. In conclusion, the fine-tuning models inherits the properties of linear probing models.

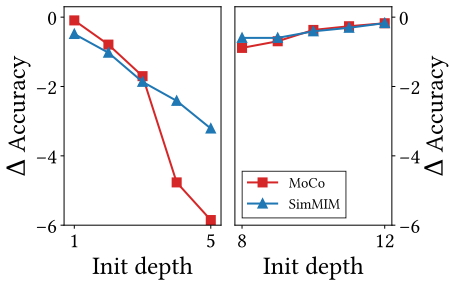
shows the predictive performance of ViT with restricted local self-attentions. As expected, the results are similar to the performance of MIM; As the kernel size decreases, the linear probing accuracy decreases but the fine-tuning accuracy increases. These results are consistent with our findings.

D Fine-Tuned Models Inherit the Properties of Pre-trained Models

CL is shape-biased and MIM is texture biased in fine-tuning scheme. In Figure 8, we demonstrate that linear probing model with CL is more shape-biased and that with MIM is texture-biased, compared with each other. As in the experiment, we calculate the classification results of ImageNet fine-tuned CL and MIM on Stylized-ImageNet, and measure the decrease in accuracy against frequency-based random noise. As we would expected, Figure C.1 shows that the property also extends to the fine-tuned model. Even though we still observe the difference between CL and MIM, the performance gap between CL and MIM is quite reduced compared to the gap between the linear probing models.

Later layers of CL and early layers of MIM are important in find-tuning phases. Figure 11 shows that later layers of CL and early layers of MIM are linearly separable. Even though this result implies that later layers of CL and early layers of MIM are important, it does not directly support that these properties are maintained in fine-tuning phases.

To demonstrate this claim in fine-tuning phases, we measure the accuracy drop of fine-tuned models using pre-trained models with a few blocks initialized. As shown in Figure D.1a, initializing a few early blocks in the pre-training models of MIM significantly harms the fine-tuning accuracy, compared with CL. This suggests that early layers of MIM play an important role in fine-tuning. Conversely, Figure D.1b shows that initializing later blocks in the pre-training models of MIM does not harms the fine-tuning accuracy significantly, suggesting that they are not important in fine-tuning compared with CL.



(a) Early layer (b) Later layer

Figure D.1: **Later layers of CL and early layers of MIM play a key role in find-tuning phases.** We measure the fine-tuning accuracy drop of pre-trained models with a few blocks initialized.

A limitation of this experiment is that the accuracy drop is evaluated in single run. Since the accuracy drop of MoCo is higher than that of SimMIM within a margin of error at the first few initialization depth in Figure D.1b, additional experiments may help improve the result. In this experiment, we use the same fine-tuning settings for MoCo and SimMIM, but experiments with tailored fine-tuning setting for each method may also be helpful.

E CL TRAINS SELF-ATTENTIONS AND MIM TRAINS MLPS

CL captures low-frequency information and MIM captures high-frequency information. Moreover, self-attentions capture low-frequency information and MLPs capture high-frequency information (Park & Kim, 2022b). Therefore, we conjecture that CL mainly trains self-attentions and MIM trains MLPs.

We measure ℓ_2 distance between initial weight and optimized weight module-wisely. Table E.1 provides the result, indicating that the weight of self-attentions changes significantly in CL learning, and in MIM learning, the weight of MLP changes dominantly.

Table E.1: **CL primarily trains self-attention (Attn) and MIM trains MLP.** We measure ℓ_2 distance between initial weight and optimized weight module-wisely. The larger the distance, the more the method trained the module.

METHOD	Attn	MLP
MoCo	**286**	338
SimMIM	247	**375**

F HYBRID MODELS OUTPERFORM CL AND MIM IN DOWNSTREAM TASKS

Section 5 demonstrates the claim—CL and MIM are complementary—only on ImageNet. To extend the effectiveness of this claim beyond ImageNet, we evaluate the pre-trained models of the hybrid models in Section 5 in another classification task and a semantic segmentation task. In particular, we measure accuracy on iNaturalist 2018 (Van Horn et al., 2018) and mIoU on ADE20K (Zhou et al., 2019). As in Table F.1, the result shows that the hybrid model of SimMIM and MoCo outperforms both SimMIM and MoCo in the downstream tasks.

Table F.1: **Hybrid models of CL and MIM outperform both CL and MIM.**

λ (IMPORTANCE OF CL)	iNat-18	ADE20k
0.0 (SimMIM)	62.1	35.4
0.2 (SimMIM + MoCo)	**68.8**	**42.2**
1.0 (MoCo)	66.2	39.7

SUPPLEMENTAL MATERIAL FOR THE DISCUSSION

Representational displacement, the distance between representations of previous and subsequent modules, may not be an appropriate metric for comparing representations of different methods. Based on Reviewer 54TF's feedback, we measure the relative displacement of representation, which is the ratio of the displacement to the norm of representation, and update the MoCo data. As shown in Figure F.1, the result shows that the relative displacement of MoCo and that of SimMIM show no significant difference except the first layer and the last few layers. Therefore, we replace the displacement experiments with the analyzes of MAE in Figure 10. As a result, we improve the consistency of Section 4; Section 4 in this revision claims that early layers of MIM and later layers of CL are important and MAE exploits such properties.

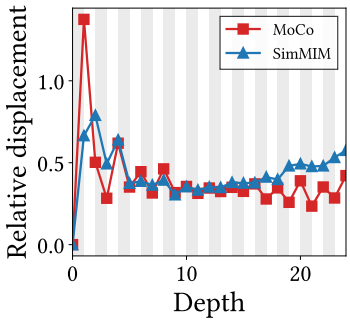


Figure F.1: **Relative displacement of CL and MIM.** We measure the relative displacement of representation, i.e., the ratio of displacement to norm of representation.