Learning with Unmasked Tokens Drives Stronger Vision Learners – Supplementary Material –

Taekyung Kim*, Sanghyuk Chun, Byeongho Heo, and Dongyoon Han*

NAVER AI Lab

Appendix

This appendix includes additional experimental analyses of our proposed method, comparing it with state-of-the-art self-supervised learning (SSL) methods and experimental results with detailed setups. We first provide the attention map visualizations in §A; we then provide 1) another applicability of our proposed method to SimMIM, 2) ablation studies, and 3) our implementation details, including hyper-parameters in §B.

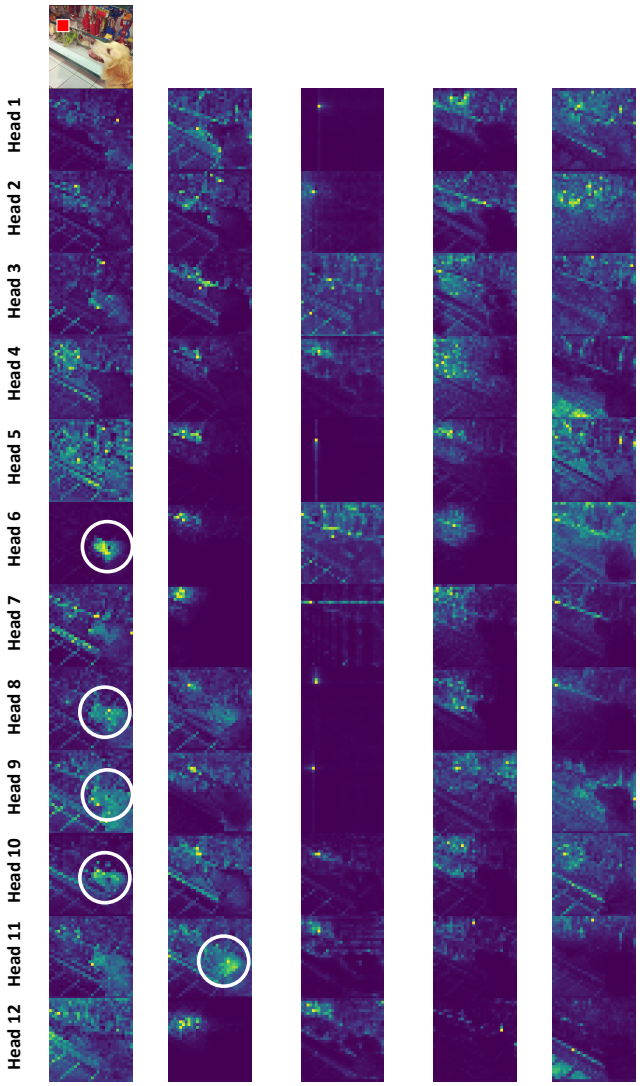
A On Distinctiveness of Attention Map

In this section, we qualitatively show the improved discriminative power of our model compared with other SSL methods [1–3, 8] and LUT through attention map visualization by visualizing all the multi-heads of the last self-attention block using sample cases. We visualize the attention maps of the entire heads of the last self-attention according to the given query patches. We compare the diverse methods to investigate the distinctive trends. Fig. A and Fig. B showcase when the input queries are from the background of the images, As shown in Fig. A, models pre-trained with DINO [2] highlight foreground regions despite the background query, which reveals DINO broadly aggregates representations across the image, losing discriminative power. Moreover, iBOT also suffers from the correlation between the representations of foreground and background patches, as observed in Fig. Ab and Fig. Bb. data2vec shows precise local discriminatibility in Fig. Ac, but indiscriminatively highlights attention in Fig. Bc. While MAE does not confuse foreground and background representations in Fig. Ad, MAE suffers another confusion in Fig. Bd, which may stem from lack of broader contexts. Besides, LUT shows enhanced discriminability between foreground and background patches in both cases.

B Experiments (cont'd)

This section presents continued experiments that further investigate the superiority and applicability of our method. We show another application of broader

* Equal contribution



(a) DINO **(b)** iBOT **(c)** data2vec **(d)** MAE **(e)** Ours

Fig. A: Attention visualization for all multi-heads of the last self-attention block. Given a sample and a query (left top on Fig A.3(a)), We visualize the attention maps of the models (with ImageNet-1K accuracies) pre-trained by DINO [2], iBOT [8], data2vec [1], MAE [3], and LUT. Each row presents the corresponding attention map of each head. White circles in the attention maps emphasize the highlighted foreground regions despite the background query. We use the ViT-B/16 architecture and a resolution of 224×224. We borrowed a sample image from `n2099601` ImageNet-1K class.

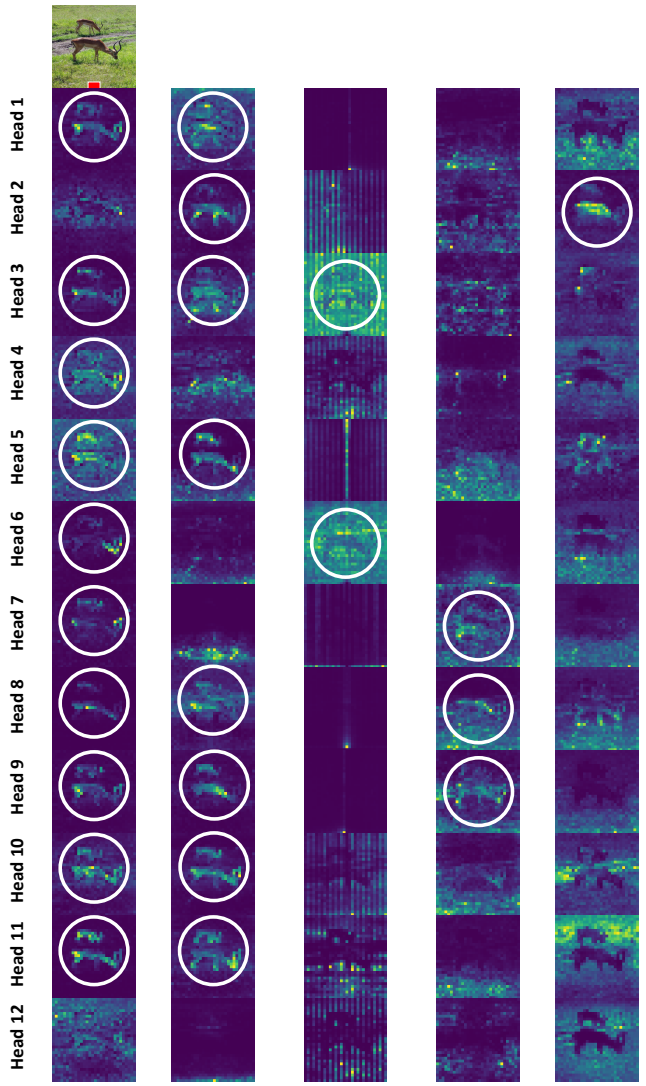


Fig. B: Attention visualization for all multi-heads of the last self-attention block. Given a sample and a query (left top on Fig A.3(a)), We visualize the attention maps of the models (with ImageNet-1K accuracies) pre-trained by DINO [2], iBOT [8], data2vec [1], MAE [3], and LUT. Each row presents the corresponding attention map of each head. White circles in the attention maps emphasize the highlighted foreground regions despite the background query. We use the ViT-B/16 architecture and a resolution of 224×224. We borrowed a sample image from `n2422699` ImageNet-1K class. The grid pattern in (c) is presumably induced by the interpolation of the relative pose bias.

Table A: Impact of broader contextualization in SimMIM. To verify the versatility of our method to other methods, we apply the proposed broader contextualized supervision to training SimMIM. All models are pre-trained and fine-tuned on ImageNet-1K. We employ ViT-B/16 trained with the image resolution of 224×224 and the identical weighting parameter of 0.25 for our context supervision loss (*i.e.*, $\mathcal{L}_{\text{BC}}$).

Method	Pre-training epochs	Accuracy (%)
SimMIM	100	81.6
LUT (SimMIM)	100	**81.8**

Table B: Loss balancing study. We study the balance A weight between global guidance and MIM loss. All the studies report fine-tuning and linear probing accuracies for each configuration which are pre-trained with ViT-B/16. All the backbones are pre-trained for 400 epochs. We mark the default settings for the study in gray.

Case	Fine-tuning (%)	Linear probing (%)
0.1	83.2	**70.7**
0.25	**83.5**	67.9
0.5	83.4	70.1
1.0	82.9	63.6

context supervision in masked image modeling beyond MAE. We finally share our experimental regimes for the ImageNet-1k fine-tuning and semantic segmentation experiments on ADE20K.

B.1 Further Applicability of Our Method

We showcase another use case of our method with another baseline. We choose a representative masked image modeling SimMIM [6]. We aim to reveal that our solution is also compatible with other masked image modeling methods that do not drop mask tokens in the encoder, such as SimMIM [6].

We pre-train the models with SimMIM, which is the baseline, and SimMIM with our method on ImageNet-1K [4] for 100 epochs and fine-tuned following the fine-turning recipe of SimMIM [6]. We primitively replace the masked image modeling part of our framework for MAE with SimMIM and employ the framework for training. As shown in Table A, our method improves SimMIM by 0.2%p despite short pre-training epochs, which shows the potential applicability of our method on MIMs.

B.2 Balancing $\mathcal{L}_{\text{BC}}$

To give a maximal impact through broader context supervision loss, we study an appropriate α in Eq. (3), and Table B shows that a loss weight of 0.25 works best, and our method's effectiveness remains up to 0.5. Moreover, though the

Table C: Training with the Broader Contextualization loss (*i.e.*, $\mathcal{L}_{\mathrm{BC}}$) only. All the models are pre-trained for 100 epochs on ImageNet-1K. Fine-tuned results on ImageNet-1K are reported.

Method	Fine-tuning (%)
Baseline	82.1
$\mathcal{L}_{\mathrm{BC}}$ only	82.0 (−0.1%p)
LUT	82.6 (+0.5%p)

highly tilted loss weights brought relatively degraded performance, these models work better than a model pre-trained by MAE.

B.3 How does training proceed when only using $\mathcal{L}_{\mathrm{BC}}$?

To further investigate its impact, we exclusively train with broader contextualization loss. We pre-train and fine-tune a ViT-B/16 on ImageNet-1K [4]. As shown in Table C, while the model pre-trained with $\mathcal{L}_{\mathrm{BC}}$ results in on par accuracy to the baseline, which suggests a broad context decent supervision to the trainable encoder. However, it decreases the accuracy 0.6%p from LUTs, demonstrating that the combination with the MIM loss learns more discriminative representations.

Table D: Hyper-parameter configurations for end-to-end fine-tuning on ImageNet-1K. All the numbers are for fine-tuning with the ImageNet-1k pre-trained backbone to the ImageNet-1K classification.

Config	Value
Optimizer	AdamW
Base learning rate	5e-4 (S), 2.5e-4 (B), 1e-3 (L)
Weight decay	0.05
Optimizer momentum	$\beta_1, \beta_2 = 0.9, 0.999$
Layer-wise learning rate decay	0.75 (S), 0.65 (B, L)
Batch size	1024
Learning rate schedule	Cosine decay
Warmup epochs	5
Training epochs	300 (S), 100 (B), 50 (L)
Resolution	224×224
Augmentation	RandAug (9, 0.5)
Label smoothing	0.1
Mixup	0.8
Cutmix	1.0
Drop path	0.1

Table E: Hyper-parameter configurations for the ADE20K finetuning. All the numbers are for transfer learning with the ImageNet-1K pre-trained backbone to the ADE20K semantic segmentation.

Config	Value
Optimizer	AdamW
Learning rate	1e-4
Weight decay	0.05
Optimizer momentum	$\beta_1, \beta_2 = 0.9, 0.999$
Layer-wise learning rate decay	0.65
Batch size	16
Learning rate schedule	Polynomial
Warmup iterations	1500
Training epochs	160k
Resolution	512×512
Drop path	0.1

B.4 Additional Implementation Details

Fine-tuning setup for ImageNet-1K classification. We list the detailed hyper-parameters for fine-tuning on ImageNet-1K [4] in Table D. Specifically, we use the AdamW optimizer and a weight decay 0.05 with a batch size of 1024. We used a layer-wise learning rate decay of 0.75 for ViT-S/16 and 0.65 for ViT-B/16 and ViT-L/16. We fine-tune ViT-S/16, ViT-B/16, and ViT-L/16 for 300, 100, and 50 epochs, respectively.

Detailed setup for ADE20K semantic segmentation. We provide the detailed hyper-parameters for transfer learning to the semantic segmentation task on ADE20K [7] in Table E. We fine-tune UperNet [5] initialized with our pre-trained model for 160k iterations with a batch size of 16. Note that we do not employ multi-scale training.

References

1. Baevski, A., Hsu, W.N., Xu, Q., Babu, A., Gu, J., Auli, M.: data2vec: A general framework for self-supervised learning in speech, vision and language. In: Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., Sabato, S. (eds.) International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 162, pp. 1298–1312. PMLR (2022) 1, 2, 3
2. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the International Conference on Computer Vision (2021) 1, 2, 3
3. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16000–16009 (2022) 1, 2, 3

4. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: Imagenet large scale visual recognition challenge. International Journal of Computer Vision **115**(3), 211–252 (2015) 4, 5, 6
5. Xiao, T., Liu, Y., Zhou, B., Jiang, Y., Sun, J.: Unified perceptual parsing for scene understanding. In: European Conference on Computer Vision. Springer (2018) 6
6. Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., Dai, Q., Hu, H.: Simmim: A simple framework for masked image modeling. In: International Conference on Computer Vision (2022) 4
7. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene parsing through ade20k dataset. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2017) 6
8. Zhou, J., Wei, C., Wang, H., Shen, W., Xie, C., Yuille, A., Kong, T.: ibot: Image bert pre-training with online tokenizer. In: International Conference on Learning Representations (2022) 1, 2, 3