Longer-range Contextualized Masked Autoencoder

Taekyung Kim* Sanghyuk Chun Byeongho Heo Dongyoon Han*
NAVER AI Lab

Abstract

Masked image modeling (MIM) has emerged as a promising self-supervised learning (SSL) strategy. The MIM pre-training facilitates learning powerful representations using an encoder-decoder framework by randomly masking some input pixels and reconstructing the masked pixels from the remaining ones. However, as the encoder is trained with partial pixels, the MIM pre-training can suffer from a low capability of understanding long-range dependency. This limitation may hinder its capability to fully understand multiple-range dependencies, resulting in narrow highlighted regions in the attention map that may incur accuracy drops. To mitigate the limitation, We propose a self-supervised learning framework, named Longer-range Contextualized Masked Autoencoder (LC-MAE). LC-MAE effectively leverages a global context understanding of visual representations while simultaneously reducing the spatial redundancy of input at the same time. Our method steers the encoder to learn from entire pixels in multiple views while also learning local representation from sparse pixels. As a result, LC-MAE learns more discriminative representations, leading to a performance improvement of achieving 84.2% top-1 accuracy with ViT-B on ImageNet-1K with 0.6%p gain. We attribute the success to the enhanced pre-training method, as evidenced by the singular value spectrum and attention analyses. Finally, LC-MAE achieves significant performance gains at the downstream semantic segmentation and fine-grained visual classification tasks; and on diverse robust evaluation metrics. Our code will be publicly available.

1. Introduction

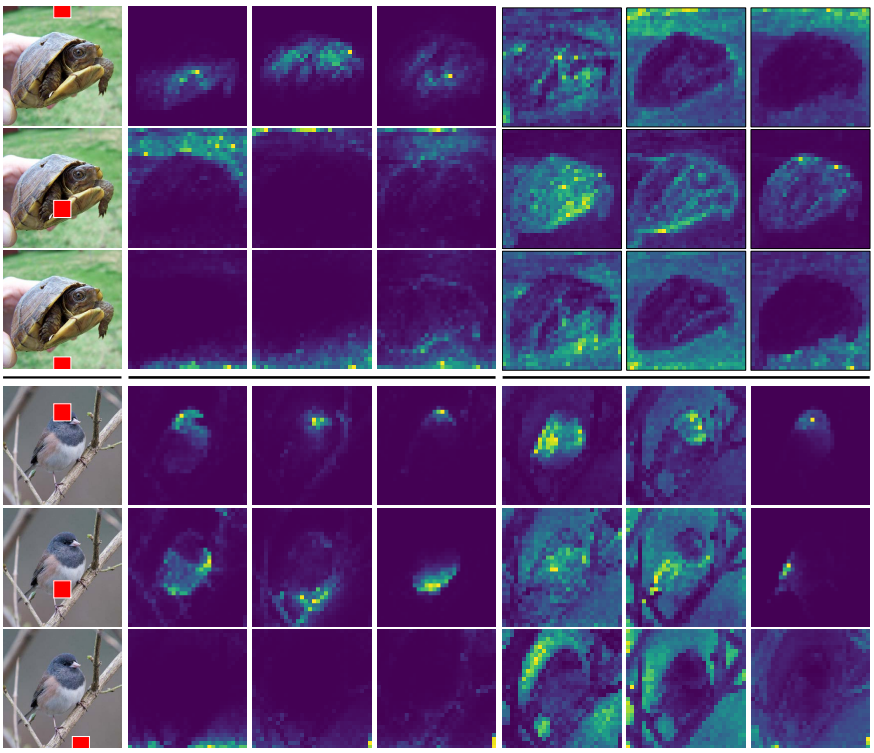
Triggered by successful transitions of Transformer [47] into vision domains [6, 16], a plethora of effective training strategies for Transformer have emerged [8, 10, 21, 43, 44]. Recent advances in masked image modeling (MIM) [3, 21, 54, 58] noticeably show great success in self-supervised learning (SSL) of Vision Transformers (ViT) by transferring the knowledge of masked language modeling [13]. Conceptually, MIM tasks consist of two parts; randomly masking out a part of inputs (*e.g.*, 75% of input pixels); and predicting the masked inputs by the decoder. This simple strategy enables a model to learn strong representations through the challenging task.

However, MIM strategies often encounter challenges such as short-range dependency on attention and limited long-rage context of the whole image. For example, Liu et al. [32] revealed that masked autoencoder (MAE) [21], a state-of-the-art MIM method, exhibits shorter average attention distances. Furthermore, we can observe that the attention pattern by MAE reveals extremely local behavior (See Fig. 1). In other words, the MAE-trained attention mechanism less integrates information across the entire image pixels and tends to focus on specific input regions (See Fig. 1b). This is presumably attributed to MIM, primarily dedicated to predicting low-level pixel details (*e.g.*, color or texture) without a comprehensive understanding of less-regional information (*e.g.*, the input structure or shape).

We aim to understand the chronic shortage in long-range dependency and how it affects MIM. We illustrate that vanilla MIM methods appear to lack longer-range dependency, while other training methods (*e.g.*, DeiT [43], MoCo v3 [10]) do not. Drawing from this, we introduce a simple solution to the short-range dependency and observe how it helps MIM to mitigate the issue. Our proposed Longer-range Contextualized Masked Autoencoder (LC-MAE) enhances the sub-optimal representation learning by offering longer-range context supervision extracting general context from the entire pixels to learn more context-generalized representations.

During training, LC-MAE minimizes the discrepancy between the encoded general context representations and the sparse representation processed by the online encoder from different views while performing MIM with a decoder. This ensures providing more contextualized visible tokens for mask tokens to attend to. The target network encodes a general representation of all pixels from a strongly augmented view to provide context information less dependent on regional changes like color distortion. In contrast, the online network encodes a sparse and unmasked view, and

*Equal contribution



(a) Input (b) MAE (c) LC-MAE (Ours)

Figure 1. **MAE lacks comprehensive region-wide attention.** We verify how attention appears differently in MAE corresponding to given queries. (a) The first column denotes the example images with different queries (red points), randomly picked patch indices. (b-c) Every set of three columns represents the highest attended maps from different heads. The turtle images have a foreground and two upper and lower background queries; the bird images have two foreground queries (upper two rows) and one background query. MAE shows localized attention maps but fails to provide comprehensive coverage of either foreground or background.

the decoder reconstructs the masked pixels using the encoded features, similar to He et al. [21]. We presume that our strategy promotes the learning of the encoder by incorporating longer-range context supervision, allowing the target network to provide a broad context for the entire pixels.

We verify the effectiveness of LC-MAE by pre-training ViT networks [16] on the ImageNet-1K benchmark [40]. Given our method's weight on improving the baseline MIM, LC-MAE-trained ViT-B/16 successfully improves linear evaluation (+2%p) and fine-tuning (+0.6%p) performance gains on ImageNet-1K over MAE. Our fine-tuning result also achieves comparable or outperformed ImageNet-1K validation accuracy (84.2%) compared with other state-of-the-art methods. LC-MAE can be transferred to the multiple fine-grained classifications and show distinguished transferability. LC-MAE further shows superior transferability and tuning robustness on INaturalist datasets. We further transfer our pre-trained model to the semantic segmentation task on ADE20K [57] and show 48.6% mIoU, a solid result in the ViT-B scale. As another benefit, LC-MAE successfully realizes robust training, which results in superior robustness results on two in-distribution benchmarks, five out-of-distribution benchmarks, and SI-Score [14].

2. Preliminary

Despite MIM's strong performance, we claim it still lacks strong attention capability after pretraining, particularly for comprehensive region-wide dependency. The upcoming spatial attention map visualizations motivate our method.

2.1. Motivation

Attention map visualizations. The attention map visualization qualitatively reveals how a model reacts to queries and reflects the learning dynamics. Fig. 1 (b) shows the attention maps concerning the given query in Fig. 1 (a) by MAE [21]. We exploit self-attention in the last block for visualization in the official ViT-B/16 MAE and visualize maps with 480×480 images from ImageNet-1K.

We observe that MAE shows narrow highlight regions for the given queries. Specifically, when a query is selected in the foreground (the 1st, 4th, and 5th rows), MAE only highlights the near patches of the given query; when a query is selected even in the background (the 2nd, 3rd, and 6th rows), we observe the same phenomenon, namely, MAE only focuses on the near patches of the given query. Based on this, we argue that MAE's attention lacks the longer-

range dependency. This may incur a lack of global understanding of the entire foreground or background concerning localizability.

2.2. Masked Image Modeling (MIM) and Beyond

We begin with a generalized formulation of MIM, addressing the limitation shown in the formulation. We then present our simple solution to remedy the limitation.

Formulation. Given an image from an augmented view u, we patchify the image into N non-overlapping patches $U = \{u_i\}_{i=1}^N$. We randomly pick masked patches $\mathcal{M}$ with a high masking ratio $r \in (0, 1)$, where $\mathcal{M} \subset \{1, 2, ..., N\}$ and $|\mathcal{M}| = \lfloor rN \rfloor$. We denote the masked image patches as $U_m = \{u_i : i \in \mathcal{M}\}$ and the remaining patches as $U_r = \{u_i : i \in \{1, 2, ..., N\}, i \notin \mathcal{M}\}$. The remaining patches are fed into the encoder f_θ and become encoded tokens $T_e = f_\theta(U_r)$. The encoded tokens are concatenated with mask tokens m_i corresponding to the positions of i-th masked patches (entire patches $U_m \cup U_r$ can be fed into the encoder [54]). The only mask tokens predict the image patches through the decoder d_ϕ. We denote i-th decoded mask token and input mask token as m_i^d and m_i where $T_d \cup \{m_i^d\}_{i\in\mathcal{M}} = d_\phi(T_e \cup \{m_i\}_{i\in\mathcal{M}})$, informally. Here, T_d is a set of decoded visible tokens. Now, the MIM pretraining objective is defined as follows:

$$\mathcal{L}_{\text{MIM}} = \sum_{i\in\mathcal{M}} ||m_i^d - u_i||_2^2, \tag{1}$$

where m_i are shared for all the positions.

Our simple solution. Eq. (1) aims to provide effective local supervision for masked tokens by reconstructing image patches via the decoder d_ϕ with limited information. However, this could lead to underutilizing complete image information (*i.e.*, longer-range contexts) due to lacking abundant remaining tokens at the same time; thus, its localizability is confined to a limited range, extending only to adjacent visual tokens from the anchor (query). We contend that it is because only reconstructed masked tokens are used for the actual loss calculation. Furthermore, the MIM loss enforces the reconstructed masked tokens to regress the corresponding target tokens in a patch-wise manner, lacking to establish strong neighboring dependencies and thereby providing inadequate supervision. We employ another loss that is expected to aid $\mathcal{L}_{\text{MIM}}$ by giving expansive supervision to visible tokens from the entire visual tokens:

$$\mathcal{L}_{\text{ours}} = \sum_{i\in\mathcal{M}} ||m_i^d - u_i||_2^2 + \alpha\mathcal{D}(T_e, g(U_m \cup U_r)), \tag{2}$$

where $\mathcal{D}(\cdot, \cdot)$ and $g(\cdot)$ denote a distance function and a global encoder. We here straightforwardly give encoded comprehensive supervision from entire tokens to visible tokens T_e. During training, the expansively supervised T_e contains extended token information so that mask tokens can leverage. This potentially gives additional localization capability beyond what the baseline possesses. The options for choosing $\mathcal{D}$ and g are indeed diverse, but we take the simplest way in the next section. Eq. (2) can involve both visible and mask tokens, but we focus on visible tokens to prevent learning collapse in mask tokens.

3. Method

In this section, we introduce Longer-range Contextualized Masked Autoencoder (LC-MAE) that addresses the short-range dependency issue in MAE.

Global contextualized supervision. The crux of our solution lies in providing a more comprehensive contextualization of entire visual tokens. Here, we opt for the elements in the newly involved loss (dubbed global guidance loss $\mathcal{L}_{\text{GG}}$) in Eq. (2). First, for the global encoder g, we implement this by simply reusing the encoder f_θ to give the supervision back to f_θ. It performs like a token-level regression between the encoders. We opt for an efficient yet strong option, momentum networks [7, 10, 18, 20]. The architecture consists of a momentum encoder and MLP head, which shares a nearly identical architecture to the online network.

Additionally, we augment the entire image patches from U to V to enhance the generalization of the encoder and avoid collapse. For the global latent features, the view v is patchified into $V = \{v_i\}_{i=1}^N$, respectively. Unlike a general MIM process, the whole patches V are encoded by the global encoder g; we denote the whole encoded tokens as $T_g^V = g(V)$, where $T_g^V = \{t_i^V\}_{i=1}^N$. Finally, the MLP head h yields global representations $\hat{v} = h(\tilde{v})$, where $\tilde{v} = \frac{1}{N}\Sigma_{i=1}^N t_i^V$ is globally pooled or each set of representations $\tilde{v} = T_g^V$. Alternatively, using aligned tokens [12] for $\hat{v}$ could benefit performance, but we simply use a pooled token.

We refer to the process as delivering *global contextualized supervision*, which involves utilizing information from entire tokens to facilitate training through comprehensively contextualized guidance.

Sparse tokens that learn broad contexts. Our encoding process obtains regional representations $T_e^{U_r} = f_\theta(U_r)$ from sparsified tokens U_r. Similar to computing global guiding representations, we aggregate the latent embeddings $u_\theta = \frac{1}{|T_e^{U_r}|}\Sigma_{t\in T_e^{U_r}} t$ through averaging. We follow the previous studies preventing training collapse by applying an MLP head h_θ to obtain $\hat{u_\theta} = h_\theta(u_\theta)$, forming architectural asymmetry to avoid collapse [9, 18]. LC-MAE can

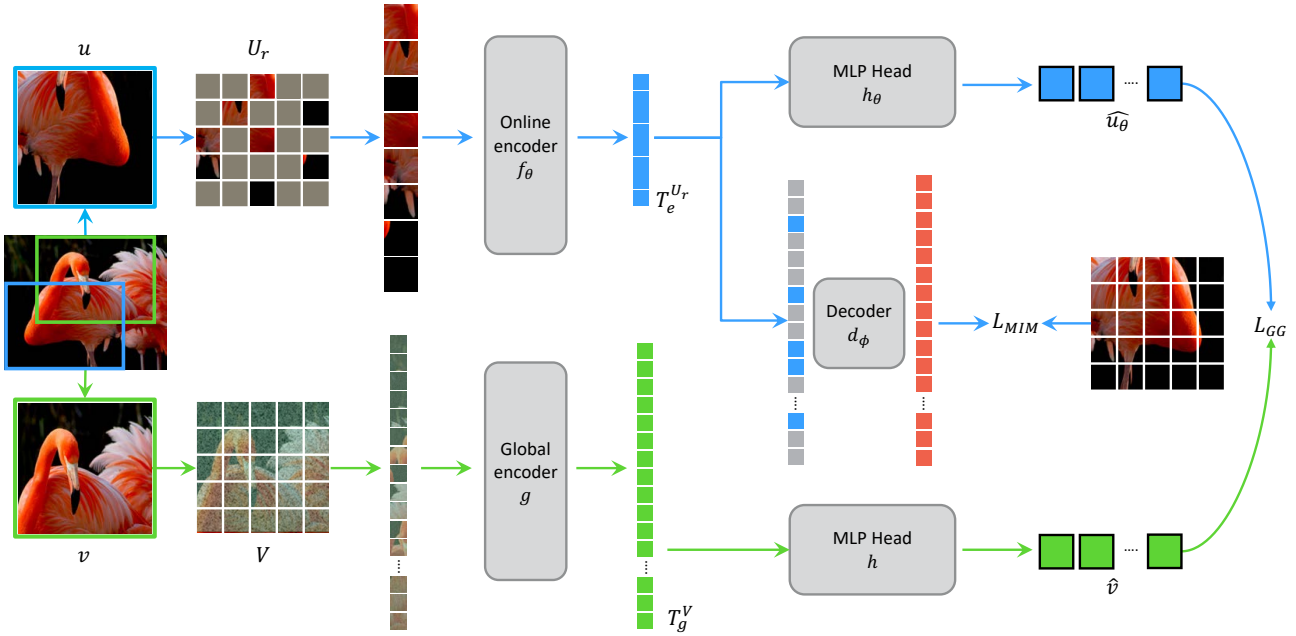


Figure 2. **Framework overview.** We introduce global contextualized guidance to enhance masked image models. Our method performs masked image modeling with undistorted sparse tokens while another global encoder guides the online network using a distorted global view. We build the networks using simple yet widely-used MLP head architectures [8–10, 18] on top of encoders to avoid collapse. We opt for a simple choice where the additional global momentum encoder mirrors the online encoder, while we may alternatively employ various options. We borrow a flamingo image from `n02007558` class in ImageNet-1K.

be interpreted as utilizing masked tokens for MIM interacting with sparse visual tokens that are employed to condense expanded context information.

On contextual discrepancies across views. We aimed to provide global contextualized supervision to visible tokens that correspond to the original view of the masked tokens. However, MIMs generally use random resized crop (RRC) [41] for giving geometric variation; we argue that using RRC may not align with our intention and could hinder learning due to divergent views providing limited shared information [9, 44]. Thus, we adopt simple resized crop (SRC) [44] instead of RRC. We conjecture the latent features from remaining sparse tokens can be more reliably guided by the semantics from the global latent. We will observe that SRC harms MAE but improves LC-MAE.

Objective function. We finalize our objective by choosing the distance function $\mathcal{D}$ in Eq. (2). We apply the normalized ℓ_2-distance for the feature distance (*i.e.*, Cosine distance). We have the aggregated global representation $\hat{v}$ and sparse one $\hat{u}_\theta$, and their ℓ_2-normalized version $\bar{u}_\theta$ and $\bar{v}$, respectively. Our global guidance loss $\mathcal{L}_{\text{GG}}$ computes the feature distance between normalized representations $\bar{u}_\theta$ and $\bar{v}$, formulated as $\mathcal{L}_{\text{GG}} = ||\bar{u}_\theta - \bar{v}||_2^2$. We conjecture that LC-MAE is agnostic to the choice of distance function since the fundamental principle of it works regardless of the distance functions, and the InfoNCE or Smoothed ℓ_1 losses also show compatibility with LC-MAE. The final objective function is:

$$\min_\theta \; \mathcal{L}_{\text{MIM}} + \alpha \mathcal{L}_{\text{GG}}, \tag{3}$$

where α controls the balance of the global guidance loss and the masked image modeling loss. The study on α gives the best fine-tuning performance with 0.25; however, it is insensitive to the choice of α; for example, α = 0.25 and 0.5 shows only 0.1% difference of fine-tuning performance using the ViT-B/16 backbone. We support all our design choices in the ablation studies in Table. 5. Our method is also applicable to SimMIM [54]-like methods with performance improvements (see Appendix for details).

3.1. Comparisons with prior arts

Prior to transitioning to our experiments, we outline the distinctions between our work and closely related studies. Several studies have been conducted recently employing multiple encoders, such as our online and target encoders. For example, a line of research excludes using additional data and employs an additional tokenizer module for the reconstruction supervision [2, 58].

Zhou et al. [58] proposed iBOT that jointly trains the target encoder and the online tokenizer. The main motivation is to align the full representations of multi-view instances among 12 different views while additionally performing masked feature reconstruction. Thus, iBOT needs

Method		Pre-training epochs (ViT-S/B/L)	Supervision	ViT-S	ViT-B	ViT-L	ADE20K
Supervised models							
DeiT [43]	ICML 2021	-	Label	79.9	81.8	-	-
DeiT-III [44]	ECCV 2022	-	Label	81.4	83.8	84.2	49.3
Cosub [45]	CVPR 2023	-	Label	81.5	**84.2**	85.3	49.3
Self-supervised models							
MoCo v3 [10]	ICCV 2021	300 / 300 / 300	Pixel	81.4	83.2	84.1	47.3
DINO [8]	ICCV 2021	800 / 800 / N/A	Pixel	81.5†	82.8†	-	46.8
iBOT [58]	ICLR 2022	3200 /1600 / 1000	Feature	**82.0**	84.0†	84.8†	**50.0**†
MAE [21]	CVPR 2022	1600 /1600 / 1600	Pixel	81.4‡	83.7‡	85.6‡	48.1
SimMIM [54]	CVPR 2022	800 / 800 / N/A	Pixel	<u>81.9</u>‡	83.8	-	-
MaskFeat [51]	CVPR 2022	N/A / 1600 / 1600	Feature	-	84.0	85.7	-
ExtreMa [52]	arXiv	300 / 300 / N/A	Feature	81.8	83.7	-	47.9
data2vec [2]	ICML 2022	800 / 800 / 1600	Feature	81.8‡	<u>84.1</u>‡	**86.6**	48.3‡
SemMAE [31]	NeurIPS 2022	N/A / 800 / N/A	Pixel	-	83.3	-	46.3
SdAE [11]	ECCV 2022	N/A / 300 / N/A	Pixel	-	84.1†	-	48.6†
MSN [1]	ECCV 2022	N/A / 600 / N/A	Feature	-	83.4	-	-
BootMAE [15]	ECCV 2022	N/A / 800 / 800	Pixel + Feature	-	**84.2**	85.9	49.1
CAN [36]	arXiv	N/A / 1600 / 800	Pixel	-	83.6	84.7	-
ConMIM [56]	ICLR 2023	300 / 800 / 1600	Dictionary	**82.0**	83.7	85.5	46.0
SIM [42]	CVPR 2023	N/A / 1600 / N/A	Feature	-	83.8	-	-
HPM [50]	CVPR 2023	N/A / 800 / 800	Pixel	-	**84.2**	85.8	48.5
LC-MAE (ours)	-	400 / 1600 / 1600	Pixel	**82.0**	**84.2**	<u>86.0</u>	<u>49.5</u>

Table 1. **Comparisons with previous models on ImageNet-1K**. We compare LC-MAE with the previous results that used vanilla Vision Transformer architectures. All models were pre-trained and fine-tuned on ImageNet-1K.We use the ViT-S/16, ViT-B/16, and ViT-L/16 architectures and a resolution of 224×224. † denotes the models pre-trained using multi-crop augmentation. ‡ denotes our reproduction results. We highlight the best numbers (in boldface) and the second-best numbers (in underlined). For a fair comparison, we did not compare methods using modules trained on extra data, such as CLIP [38] or VQGAN [17].

multi-crops varying in diverse scales and augmentations. iBOT involves only mask tokens to learn target information, which incurs learning partial information, but we nevertheless speculate that leveraging multi-crops diminishes this issue. In contrast, our aim is to employ visible tokens to reinforce the understanding of longer-range context by offering complete information from a single view.

Baevski et al. [2] proposed data2vec that performs patch-wise feature prediction via masked tokens. Despite the target features being generated from entire images, data2vec may implicitly guide MIM with the global context. Specifically, we presume only mask tokens contribute to regressing the context supervision, so the interaction between mask tokens and visible tokens lacks utilizing the given contextualized information. Therefore, the patch-wise regression to the token representations may not adequately establish strong neighboring dependencies. We conjecture this eventually leads to inferior localization performance.

4. Experiment

In this section, we demonstrate our method by pre-training and fine-tuning on ImageNet-1K and conduct extensive comparisons with state-of-the-art methods. We further transfer our models to the ADE20K segmentation and various downstream datasets to confirm transferability.

4.1. ImageNet-1K Classification

Architecture. We use the standard Vision Transformer (ViT) [16] with a patch size of 16 for all experiments (*i.e.*, ViT-B/16) to fairly compare with prior arts. We use the 8-layer transformer decoder [21] for masked image modeling. We further adopt online and global MLP heads on the top of the encoders to aggregate global context from representations; each consists of two and four fully-connected layers with the embedding dimension of 4096, batch normalization layers [25], and ReLUs [30] following the previous methods [8, 10, 18]. Note that LC-MAE works even with symmetric heads. All the decoder and MLP heads are only used during training.

Pre-training setup. We follow the identical ImageNet-1K [40] pre-training protocol[1] [21]. Our model is pre-trained for 1600 epochs with 40 warmup epochs, batch

[1]We use the publicly available codebase in https://github.com/facebookresearch/mae

size of 4096, and input resolution of 224×224. We use AdamW [34] with momentum (0.9, 0.999). The learning rate is set to 1.5×10^{-4} with cosine learning rate decay [33]. We adopt a layer-wise learning rate decay of 0.65. We set a mask ratio for Eq. (1) to 0.75, a momentum decay rate τ in the target network to 0.996, and the weight of our global guidance loss (α in Eq. (3)) to 1.0 and 0.25 for the ViT-S/16 and ViT-B/16 architectures, respectively. We employ the simple resized crop [44] for geometric augmentation, color jittering, and the three augment [43] consists of Gaussian blur, grayscale, and solarization. All models are pre-trained using 8 V100-32GB GPUs.

Results. We compare our method with previous SSL methods [1, 2, 8, 10, 11, 15, 21, 31, 36, 42, 50–52, 54, 56, 58]. Table 1 shows the evaluation results on the ViT-S/B/L backbones. Our LC-MAE achieves an 82.0%, 84.2%, and 86.0% top-1 accuracy on ViT-S/16, ViT-B/16, and ViT-L/16, which improves 0.6%p, 0.6%p, and 0.4%p over the baseline, respectively. Moreover, LC-MAE outperforms other self-supervised learning methods by a large margin except for some masked feature models. This comes to a head with a smaller ViT-S/16, where most of the results are saturated, but this is presumably due to the low capability of the backbone and the high flexibility of masked feature models. LC-MAE would take advantage of further improvements using masked feature models as the baseline. The results highlight the efficacy of our proposed global contextualized supervision in enhancing MIM, which showcases its significant potential for further improvements.

Computational costs. Our method includes extra computation from forward inference with images, so there is a slight increase in computational demands. However, our method achieves a top-1 accuracy of 83.6% at 400 epochs, which matches MAE's accuracy at 1600 epochs, despite our significantly shorter GPU wall time. Specifically, our method takes 119 hours to complete 400 epochs of training, which is roughly half the training time of MAE's 1600 epochs, which requires 223 hours.

4.2. ADE20K Semantic Segmentation

To validate the transferability of our pre-trained model to dense prediction tasks, we evaluate semantic segmentation performances on ADE20K [57]. We follow the standard training protocol [21]; the models are fine-tuned for 160K iterations using UperNet [53] with a batch size of 16 and a resolution of 512×512. Other detailed hyper-parameters for training are listed in Appendix. The rightmost column in Table 1 shows the mIoU performance comparison. LC-MAE also outperforms the competing methods, including SSL and supervised learning methods. This outcome can be attributed to the improved dense prediction capability.

Method	iNat 2018	iNat 2019	iNat 2021-mini
BYOL	69.8 (68.6±0.9)	77.4 (76.7±0.8)	70.5 (69.1±1.1)
MoCo v3	70.1 (69.4±0.5)	77.6 (77.2±0.4)	70.9 (70.5±0.5)
DINO†	72.1 (71.9±0.2)	79.4 (79.0±0.4)	73.0 (72.8±0.1)
iBOT†	73.8 (73.5±0.2)	79.9 (79.5±0.4)	74.5 (74.4±0.1)
data2vec	75.2 (74.5±0.7)	80.6 (80.0±0.5)	76.2 (75.5±0.9)
MAE	74.6 (74.5±0.1)	80.2 (80.0±0.1)	75.7 (75.5±0.2)
LC-MAE	**75.8 (75.3±0.3)**	**81.0 (80.5±0.4)**	**76.7 (76.3±0.3)**

Table 2. **Transfer learning results on iNaturalists**. We further present the end-to-end fine-tuning accuracies on the iNaturalist 2018, iNaturalist 2019, and mini iNaturalist 2021 datasets [46]. We report the best results along with the mean ± std of the set of accuracies obtained from grid searches for each method. † denotes the models pre-trained using multi-crop augmentations. Our method consistently outperforms the competitors in terms of the best accuracies, further showcasing remarkable tuning robustness.

4.3. Transfer Learning

iNaturalist datasets. To further compare the transferability of learned representations, we measure image classification accuracies by fine-tuning the ImageNet-1K pre-trained models on iNaturalist 2018, iNaturalist 2019, and mini iNaturalist 2021 [46], which are highly imbalanced with different number of images per class. We compare LC-MAE with MoCo v3 [10], BYOL [18], DINO [8], iBOT [58], and MAE [21]. All the models are pre-trained ViT-B/16 with a resolution of 224×224. We report the maximum accuracy and the mean and standard deviation of the accuracies obtained by grid searches of learning rates and weight decay, following the protocol [28]. Table 2 shows LC-MAE outperforms the competitors across all datasets, which reveals superior transferability; moreover, our model benefits tuning robustness.

Fine-Grained Visual Classification (FGVC) datasets. We further validate fine-tuning classification accuracies on CIFAR-10 [29], CIFAR-100 [29], CUB-200 [48], Aircraft [35], Birds [24], Flowers [37], and Dogs [27] following the same evaluation protocol as above. Table 3 showcases LC-MAE achieves the best number on average and outstanding numbers overall, which shows improved transferability and tuning robustness across datasets again.

5. Analysis and Discussion

Here we provide ablation studies and analyses to give some intuitions from how global contextualized supervision actually works through singular value spectrums and robustness evaluations.

5.1. Robustness Evaluation

We evaluate the robustness of various methods, including DINO [8], iBOT [58], MAE [21], and data2vec [2] with

Method	Aircraft	Birds	CUB-200	CIFAR-10	CIFAR-100	Dogs	Flowers	Average
DINO†	87.0 (86.0±0.6)	83.9 (83.4±0.5)	85.1 (84.9±0.3)	99.0 (98.9±0.1)	91.3 (90.7±0.5)	84.8 (84.6±0.3)	98.8 (98.7±0.1)	90.0
iBOT†	87.3 (86.7±0.6)	85.5 (85.1±0.5)	85.9 (85.5±0.3)	**99.2** (98.8±0.6)	**92.0 (91.1±0.9)**	86.0 (85.7±0.3)	**99.0 (99.0±0.1)**	90.7
MAE	88.1 (87.3±0.9)	84.2 (84.0±0.3)	84.6 (84.3±0.2)	98.8 (98.7±0.1)	90.0 (89.7±0.3)	86.8 (86.4±0.3)	98.1 (97.8±0.3)	90.1
data2vec	87.3 (86.6±0.7)	84.1 (83.5±0.5)	84.4 (83.9±0.4)	98.8 (98.7±0.1)	91.2 (91.0±0.2)	85.7 (85.3±0.3)	96.7 (94.4±3.3)	89.7
LC-MAE	**89.2 (88.3±0.9)**	**86.0 (85.3±0.6)**	**86.5 (85.7±0.6)**	99.1 **(98.9±0.1)**	91.0 (90.7±0.4)	**87.4 (86.7±0.5)**	98.4 (98.2±0.2)	**91.1**

Table 3. **Transfer learning results.** We present the end-to-end fine-tuning accuracies on multiple datasets, reporting the best results along with the mean ± std of the accuracies from grid searches. Our method mostly outperforms the competitors at the best accuracies, further showcasing the robustness among different training hyper-parameters. † denotes the models pre-trained using multi-crop augmentation.

	IN-1k↑	IN-V2↑	IN-Real↑	IN-A↑	IN-O↑	Sketch↑	IN-R↑	Cocc↑	ObjNet↑	SI-size↑	SI-loc↑	SI-rot↑
DINO†	83.1	72.8	87.6	36.3	60.7	35.7	48.2	77.8	36.4	57.8	37.0	43.8
IBOT†	83.5	73.5	87.9	39.4	62.0	37.8	50.2	78.6	37.1	58.2	37.6	43.9
MAE	83.7	72.9	88.2	36.7	**65.4**	35.9	48.9	78.4	37.6	58.0	38.7	42.7
data2vec	84.1	74.2	88.5	41.6	62.2	**38.7**	**53.0**	79.1	**40.3**	57.9	38.6	43.8
LC-MAE	**84.2**	**74.2**	**88.6**	**42.5**	64.1	38.2	52.1	**79.2**	38.9	**59.8**	**40.7**	**44.9**

Table 4. **Robustness evaluation.** We evaluate the robustness of the ImageNet-1K-pretrained representative methods: DINO, iBOT, MAE, and data2vec with our LC-MAE on in-distribution generalization (IN-V2/Real) and out-of-distribution (IN-A/IN-O/Sketch/R/Cocc/Obj) benchmarks. We also evaluate the capability to detect spurious correlations with background on SI-Score metrics [14]. We highlight the best numbers (in boldface) and the second-best numbers (in underlined). † denotes the models pre-trained using multi-crop augmentation. Ours surpasses others significantly than indicated by the ImageNet-1K numbers, particularly more on localization-related metrics.

LC-MAE on various robustness benchmark. We verify how our method impacts model robustness. We employ two in-distribution benchmarks including ImageNet-V2 [39] and ImageNet-Real [5]) and four out-of-distribution benchmarks ImageNet-A [23], ImageNet-O [23], ImageNet-R [22], ImageNet-Sketch [49], and ObjectNet [4]. We further use SI-Score [14] to test spurious correlations with the background. Lastly, we evaluate the center occlusion benchmark that zeroes the center patch in the ImageNet-1K evaluation images. As shown in Table 4, LC-MAE achieves outstanding performance on all the benchmarks.

5.2. Ablation Study

Here, we conduct ablation studies of LC-MAE pre-training under various available configurations. We select ViT-B/16 as the base model and train it for 400 epochs on ImageNet-1K as the fixed pre-training setup. Each model is then individually pre-trained. We report the top-1 fine-tuning and linear probing accuracies for each study.

Loss function. We first explore various losses for the global guidance loss in Table 5a. While all objectives yield considerable performance as expected above, the cosine distance of latent representations of global and partial information works best when pre-training by LC-MAE.

Type of global supervision. We study the effectiveness of various guidance approaches in Table 5b. We mainly compare token-wise supervision and globally aggregated supervision. While all the types yield performance gains, the global guidance works the best, improving 0.7%p in fine-tuning even only with 400 epochs. The global guidance outperforms the combination of token-wise and global guidance, implying that the additional token-wise guidance may conflict with the global one, which is presumably due to the alignment between the set of tokens.

Masking ratio at target encoder. We argue the information in the target latent representations should remain globally. Table 5c shows that the model without masking outperforms all the counterparts.Moreover, the fine-tuning accuracy of the models with masking even underperforms the baseline, implying that transferring coarse information carelessly may harm the capability of learning representation.

Tokens for global guidance. We mainly use the visible tokens for giving guidance but study whether CLS-token can be an alternative in Table 5d. We observe using visible tokens is preferred for LC-MAE. Considering latent features undergo masked auto-encoding, these results imply that explicitly using global information is more effective than using implicit information via CLS-token.

Guided tokens. We investigate which tokens should learn the guiding information, considering both visible and mask tokens. While we designed with visible tokens, Table 5e illustrates that training solely with visible tokens yields a superior outcome, aligning with our previous expectation.

Image crop type. This study highlights how performance is affected by the disparity between the two views in our method. There would be many comparing options, we choose Random resized crop (RRC) [41] and simple resized

Case	ft	lin
None	82.8	61.5
InfoNCE	83.0	66.7
Smoothed ℓ_1	83.2	60.2
Cosine distance (Cos)	**83.5**	**67.9**

(a) **Loss function**. "Cos" works the best.

Case	ft	lin
None	82.8	61.5
Token-wise guidance	83.2	64.1
Global guidance	**83.5**	**67.9**
Token-wise + Global	83.1	66.6

(b) **Type of guidance**. Our choice beats others.

Case	Mask ratio	ft	lin
None	-	82.8	61.5
Global guidance	0	**83.5**	**67.9**
Global guidance	0.5	82.6	63.2
Global guidance	0.75	82.5	63.8

(c) **Target's masking ratio**. Global encoder needs noncorrupted views.

Case	ft	lin
None	82.8	61.5
w/ CLS token	83.2	**71.0**
w/ visual tokens (VTs)	**83.5**	67.9

(d) **Tokens for guidance**. Using visual tokens works better.

Case	ft	lin
None	82.8	61.5
w/ visible and mask tokens	83.0	67.2
w/ visible tokens only	**83.5**	**67.9**

(e) **Guided tokens**. Guiding visible tokens performs better.

Method	Epochs	Views RRC	Views SRC	ft	lin
MAE	400	✓		82.8	61.5
MAE	400		✓	82.5	64.2
LC-MAE	400	✓		83.4	67.1
LC-MAE	400		✓	**83.5**	**67.9**

(f) **View discrepancy**. Ours benefits reduced differences among views.

Table 5. **Ablation studies**. All the studies report fine-tuning (ft) and linear probing (lin) 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.

crop (SRC) [44] for comparison. Table 5f shows the model pre-trained with SRC exceeds the fine-tuning accuracy of the case of RRC. Since RRC is more compatible with MAE than SRC, performance improvements are not observed in MAE. Our method benefits from SRC, which indicates that the global information needs to align closely with the view of the online encoder, thereby facilitating training.

5.3. Spectral Analysis

We provide additional analysis on the learned layer-wise representations LC-MAE and the baseline. Inspired by the previous studies [26], we measure the singular values (SVs) of the covariance of features, *i.e.*, how the features are spread in the embedding space. More specifically, we compute a feature covariance matrix on ImageNet-1K validation set (*i.e.*, the covariance matrix has a shape of 50k×50k), and compute the SVs of the covariance matrix. Fig. 3 shows a spectrum of log of singular value gaps between MAE and LC-MAE across the layers. The singular values of LC-MAE surpass the values of MAE across the rank indices in the last layers, while both methods have similar singular values on earlier layers. The results reveal that LC-MAE have larger singular values at the output-side layers, indicating a higher rank of the feature space [19, 55]. In other words, LC-MAE utilizes the output feature space better than MAE, owing to the global understanding prompted by global guidance.

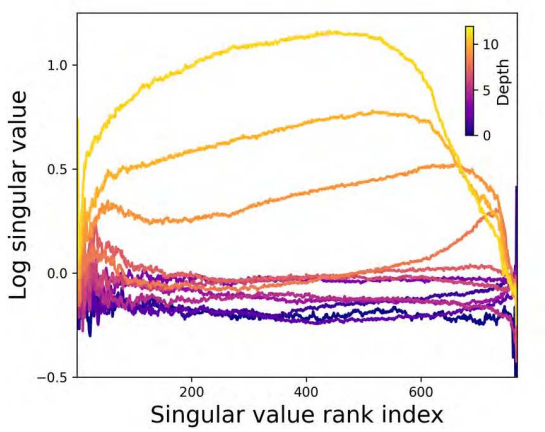


Figure 3. **Singular value (SV) spectrums**. We plot the difference of SVs between the baseline and LC-MAE for each layer, showing large gaps ($\geq$0), particularly for the later layers closer to the output.

6. Conclusion

We have introduced a novel framework to address the limited global understanding of images inherent in masked autoencoders (MAE). We have argued MAE holds short-range dependency due to lacking a comprehensive understanding of entire pixels. By visualizing attention maps, we have shown that MAE exhibits incomplete coverage of foreground or background regions. We conjectured this is caused by the potential absence of global context in learned visible tokens when interacting with mask tokens in self-attentions. Based on the observation, we have proposed LC-MAE pre-training method, minimizing the discrepancy between the global features and sparse visual tokens through our global guidance loss. The global contextualized supervision enhanced MAE by a large margin on ImageNet-1K and ADE20K, and LC-MAE significantly outperformed other state-of-the-art competitors. LC-MAE further offers significant improvements in transfer learnings, including the iNatrualist and FGVC datasets. Finally, our analyses with robust evaluations and spectral analysis demonstrated that LC-MAE can serve as a simple yet effective supplement for masked image modeling.

Appendix

This supplementary material 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; we then give another applicability of our proposed method, an extra study on balancing global guidance, and additional implementation details.

A. Further Analysis

In this section, we qualitatively show the improved discriminative power of our model compared with other SSL methods [2, 8, 21, 58] and LC-MAE through attention map visualization by visualizing all the multi-heads of the last self-attention block using sample cases.

A.1. On Discriminative of Attention Map

We further visualize the attention maps of the entire heads of the last self-attention according to 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 [8] 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 discriminability 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 global understanding. Besides, LC-MAE 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 global guidance 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. Applicablity of Global Guidance

We showcase another use case of our global guidance with another baseline. We chose a representative masked image modeling SimMIM [54]. Our aim is to reveal that our solution is also compatible with other masked image modeling methods that do not drop mask tokens in the encoder.

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

Table A. **Impact of global guidance in SimMIM**. To verify the versatility of our method to other methods, we apply the proposed global 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 the global guidance loss (*i.e.*, $\mathcal{L}_{\text{GG}}$).

Case	ft	lin
0.1	83.2	**70.7**
0.25	**83.5**	67.9
0.5	83.4	70.1
1.0	82.9	63.6

Table B. **Loss balancing**. We study the balance A weight between global guidance and MIM loss. All the studies report fine-tuning (ft) and linear probing (lin) 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.

We pre-train the models with SimMIM, which is the baseline, and SimMim with our method on ImageNet-1K [40] for 100 epochs and fine-tuned following the fine-turning recipe of SimMIM [54]. We primitively replace the masked image modeling part of our framework for MAE with SimMIM. 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 global guidance

To give a maximal impact through global guidance 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 highly tilted loss weights brought relatively degraded performance, these models work better than a model pre-trained by MAE.

B.3. Additional Implementation Details

Fine-tuning setup for ImageNet-1K classification. We list the detailed hyper-parameters for fine-tuning on ImageNet-1K [40] in Table C. Specifically, we use the AdamW optimizer and a weight decay of 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 learn-

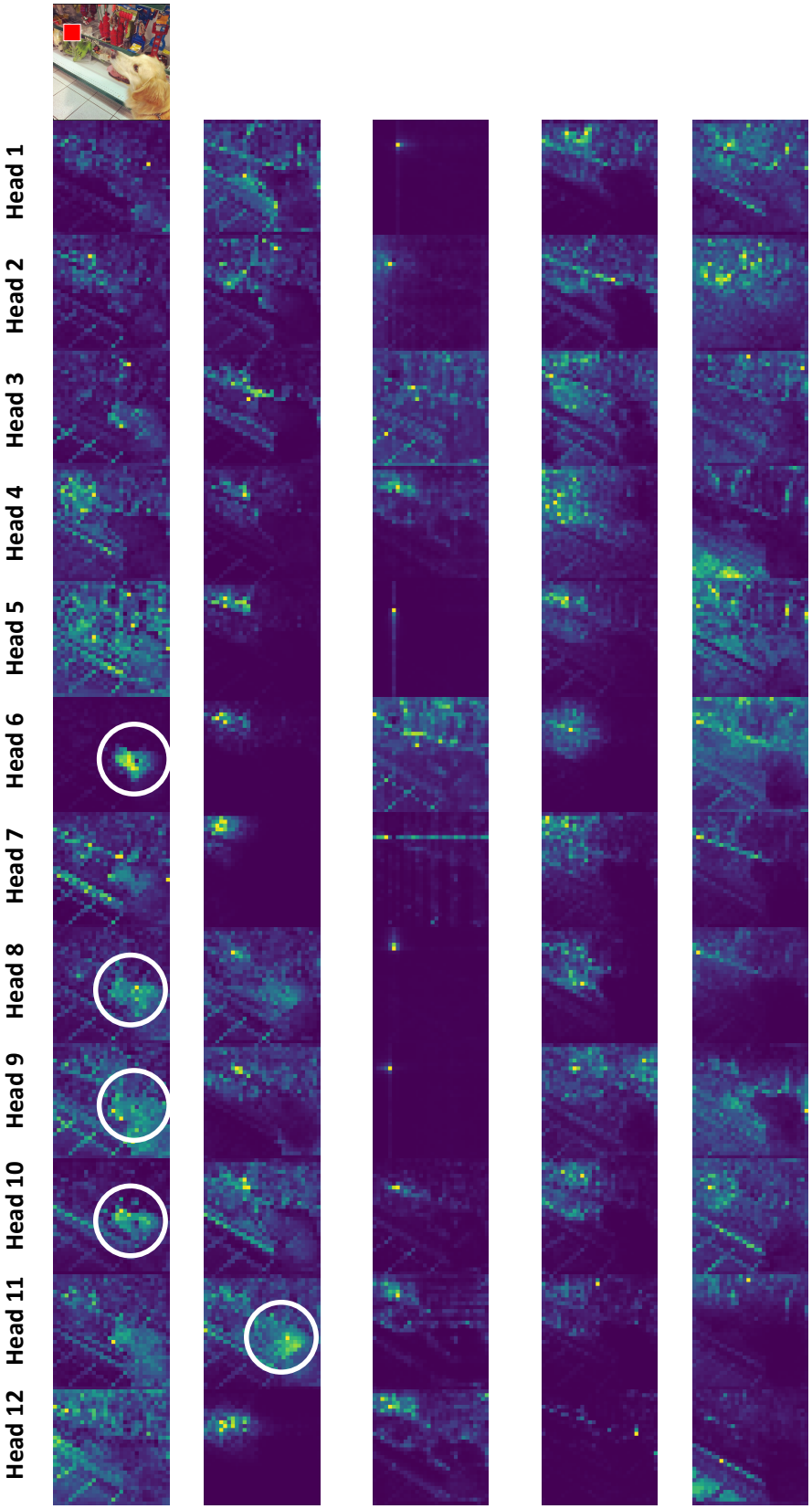


Figure A. **Attention visualization for all multi-heads** of the last self-attention block. Given a sample and a query (left top on Fig A.3(a)), We visualize the attention maps of the models (with ImageNet-1K accuracies) pre-trained by DINO [8], iBOT [58], , data2vec [2], MAE [21], and LC-MAE. Each row 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.

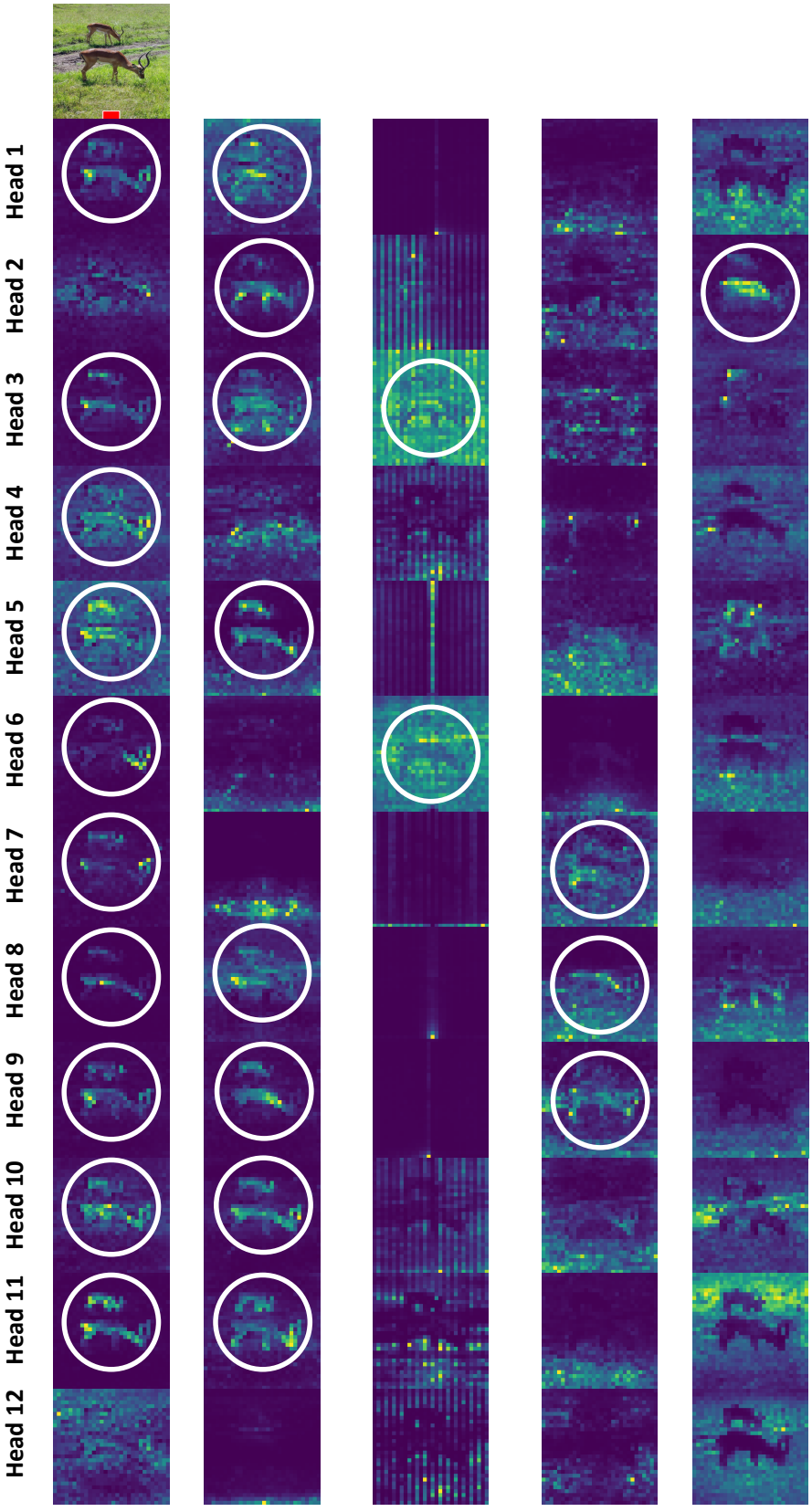


Figure B. **Attention visualization for all multi-heads** of the last self-attention block. Given a sample and a query (left top on Fig A.3(a)), We visualize the attention maps of the models (with ImageNet-1K accuracies) pre-trained by DINO [8], iBOT [58], data2vec [2], MAE [21], and LC-MAE. Each row 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.

ing to the semantic segmentation task on ADE20K [57] in Table D. We fine-tune UperNet [53] initialized with our pre-trained model for 160k iterations with a batch size of 16. Note that we do not employ multi-scale training and testing.

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 C. **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
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

Table D. **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.

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. *arXiv preprint arXiv:2204.07141*, 2022. 5, 6

[2] Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. data2vec: A general framework for self-supervised learning in speech, vision and language. In *International Conference on Machine Learning*, pages 1298–1312. PMLR, 2022. 4, 5, 6, 9, 10, 11

[3] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. In *International Conference on Learning Representations*, 2021. 1

[4] Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. *Advances in Neural Information Processing Systems*, 2019. 7

[5] Lucas Beyer, Olivier J Hénaff, Alexander Kolesnikov, Xiaohua Zhai, and Aäron van den Oord. Are we done with imagenet? *arXiv preprint arXiv:2006.07159*, 2020. 7

[6] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020. 1

[7] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *Proceedings of Advances in Neural Information Processing Systems*, 2020. 3

[8] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision*, 2021. 1, 4, 5, 6, 9, 10, 11

[9] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15750–15758, 2021. 3, 4

[10] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. *arXiv preprint arXiv:2104.02057*, 2021. 1, 3, 4, 5, 6

[11] Yabo Chen, Yuchen Liu, Dongsheng Jiang, Xiaopeng Zhang, Wenrui Dai, Hongkai Xiong, and Qi Tian. Sdae: Self-distillated masked autoencoder. In *European Conference on Computer Vision*, pages 108–124. Springer, 2022. 5, 6

[12] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, 2013. 3

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 1

[14] Josip Djolonga, Jessica Yung, Michael Tschannen, Rob Romijnders, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Matthias Minderer, Alexander D'Amour, Dan Moldovan, et al. On robustness and transferability of convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16458–16468, 2021. 2, 7

[15] 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. *arXiv preprint arXiv:2207.07116*, 2022. 5, 6

[16] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 1, 2, 5

[17] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12873–12883, 2021. 5

[18] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, koray Kavukcuoglu, Remi Munos, and Michal Valko. Bootstrap Your Own Latent - A New Approach to Self-Supervised Learning. In *Advances in Neural Information Processing Systems*, pages 21271–21284. Curran Associates, Inc., 2020. 3, 4, 5, 6

[19] Dongyoon Han, Sangdoo Yun, Byeongho Heo, and YoungJoon Yoo. Rethinking channel dimensions for efficient model design. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 732–741, 2021. 8

[20] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. *arXiv preprint arXiv:1911.05722*, 2019. 3

[21] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16000–16009, 2022. 1, 2, 5, 6, 9, 10, 11

[22] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8340–8349, 2021. 7

[23] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15262–15271, 2021. 7

[24] Grant Van Horn, Steve Branson, Ryan Farrell, Scott Haber, Jessie Barry, Panagiotis G. Ipeirotis, Pietro Perona, and Serge J. Belongie. Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 595–604, 2015. 6

[25] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, pages 448–456. pmlr, 2015. 5

[26] Li Jing, Pascal Vincent, Yann LeCun, and Yuandong Tian. Understanding dimensional collapse in contrastive self-supervised learning. *arXiv preprint arXiv:2110.09348*, 2021. 8

[27] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Li Fei-Fei. Novel dataset for fine-grained image categorization. In *First Workshop on Fine-Grained Visual Categorization, IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Colorado Springs, CO, 2011. 6

[28] Simon Kornblith, Jonathon Shlens, and Quoc V Le. Do better imagenet models transfer better? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2661–2671, 2019. 6

[29] Alex Krizhevsky. Learning multiple layers of features from tiny images. In *Tech Report*, 2009. 6

[30] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017. 5

[31] Gang Li, Heliang Zheng, Daqing Liu, Chaoyue Wang, Bing Su, and Changwen Zheng. Semmae: Semantic-guided masking for learning masked autoencoders. *arXiv preprint arXiv:2206.10207*, 2022. 5, 6

[32] Xingbin Liu, Jinghao Zhou, Tao Kong, Xianming Lin, and Rongrong Ji. Exploring target representations for masked autoencoders. *arXiv preprint arXiv:2209.03917*, 2022. 1

[33] Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*, 2017. 6

[34] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. 6

[35] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 6

[36] Shlok Mishra, Joshua Robinson, Huiwen Chang, David Jacobs, Aaron Sarna, Aaron Maschinot, and Dilip Krishnan. A simple, efficient and scalable contrastive masked autoencoder for learning visual representations. *arXiv preprint arXiv:2210.16870*, 2022. 5, 6

[37] M-E. Nilsback and A. Zisserman. Automated flower classification over a large number of classes. In *Proceedings of the Indian Conference on Computer Vision, Graphics and Image Processing*, 2008. 6

[38] Alec Radford, Jong Wook Kim, Chris Hallacy, A. Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 5

[39] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, 2019. 7

[40] 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*, 115(3):211–252, 2015. 2, 5, 9

[41] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, 2015. 4, 7

[42] Chenxin Tao, Xizhou Zhu, Weijie Su, Gao Huang, Bin Li, Jie Zhou, Yu Qiao, Xiaogang Wang, and Jifeng Dai. Siamese image modeling for self-supervised vision representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2132–2141, 2023. 5, 6

[43] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. Training data-efficient image transformers &distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357, 2021. 1, 5, 6

[44] Hugo Touvron, Matthieu Cord, and Hervé Jégou. Deit iii: Revenge of the vit. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIV*, pages 516–533. Springer, 2022. 1, 4, 5, 6, 8

[45] Hugo Touvron, Matthieu Cord, Maxime Oquab, Piotr Bojanowski, Jakob Verbeek, and Hervé Jégou. Co-training 2l submodels for visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11701–11710, 2023. 5

[46] 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/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8769–8778, 2018. 6

[47] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017. 1

[48] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The caltech-ucsd birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011. 6

[49] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems*, 2019. 7

[50] Haochen Wang, Kaiyou Song, Junsong Fan, Yuxi Wang, Jin Xie, and Zhaoxiang Zhang. Hard patches mining for masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (CVPR)*, 2023. 5, 6

[51] 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/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 5

[52] Zhirong Wu, Zihang Lai, Xiao Sun, and Stephen Lin. Extreme masking for learning instance and distributed visual representations. *arXiv preprint arXiv:2206.04667*, 2022. 5, 6

[53] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *European Conference on Computer Vision*. Springer, 2018. 6, 12

[54] 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 *International Conference on Computer Vision*, 2022. 1, 3, 4, 5, 6, 9

[55] Zhilin Yang, Zihang Dai, Ruslan Salakhutdinov, and William W Cohen. Breaking the softmax bottleneck: A high-rank rnn language model. In *International Conference on Learning Representations*, 2018. 8

[56] Kun Yi, Yixiao Ge, Xiaotong Li, Shusheng Yang, Dian Li, Jianping Wu, Ying Shan, and Xiaohu Qie. Masked image modeling with denoising contrast. *International Conference on Learning Representations*, 2023. 5, 6

[57] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2, 6, 12

[58] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. In *International Conference on Learning Representations*, 2022. 1, 4, 5, 6, 9, 10, 11