SAM Helps SSL: Mask-guided Attention Bias for Self-supervised Learning

Kensuke Taguchi*[1]
kensuke.taguchi.xm@kyocera.jp
Takehiko Kawai*[1]
takehiko.kawai.yb@kyocera.jp
Wataru Imaeda[1]
wataru.imaeda.xm@kyocera.jp
Hironobu Fujiyoshi[2]
fujiyoshi@isc.chubu.ac.jp

[1] KYOCERA Corporation, Japan
[2] Chubu University, Japan
*Equal contribution

Abstract

The vision transformer (ViT) and self-supervised learning (SSL) are key technologies for accelerating data scalability, contributing to the emergence of a foundation model in computer vision. In this paper, we focus on the potential of masks generated by the Segment Anything Model (SAM), a foundation model for image segmentation, and propose a novel method for SSL, named "mask-guided attention bias". Mask-guided attention bias is designed to encode SAM-generated masks, which are spatially and semantically decomposed information about an image. It is applied to the self-attention of ViT as guidance for an SSL process. Since self-attention can capture a wide range of spatial dependencies, mask-guided attention bias effectively adds spatial and semantic guidance to various forms of SSL, thus improving the decodability and labeling efficiency of SSL representations. We show that our method improves the accuracy of linear probing, few-shot learning, and fine-tuning in general. In particular, our method achieves 81.3% linear probing accuracy (outperforming vanilla MAE by 3.2%) and 89.5% fine-tune accuracy (outperforming vanilla DINO by 0.4%) on ImageNet100.

1 Introduction

Transformers [35] have received much attention due to their success in natural language processing (NLP) [5, 11, 27, 28]. In the field of computer vision, transformers are applied as the vision transformer (ViT) [12] and introduced as an alternative architecture to convolutional neural networks (CNNs) [21]. ViT models trained on large amounts of data outperform CNNs in many tasks [12, 37] and are scalable according to the amount of training data. Because the scalability of ViT means that it consumes large amounts of data [34], the pretraining process is critical to improve the data efficiency of downstream tasks. In general, labeling costs are an important issue because effective pretraining processes require large amounts of data. Self-supervised learning (SSL), which allows pretraining without supervised labels, is a recent breakthrough technology for mitigating supervised labeling costs [7, 18, 25].

The success of SSL has contributed to the emergence of foundation models [4, 33]. A foundation model is a model that has been trained on a broad set of data so that it can be adapted to a wide range of downstream tasks. For example, CLIP [29] is a vision-language foundation model trained on 400 million text–image pairs. Since CLIP can extract features from images and text into a common multimodal representation space, it has strong zero-shot generalization with text prompts. The Segment Anything Model (SAM) [20] is a foundation model for image segmentation tasks. SAM was trained on a broad dataset to be promptable, allowing transfer of zero-shot prompts to new image distributions and tasks for 2D images.

Inspired by the capability of foundation models, some SSL methods incorporate a vision–language foundation model like CLIP to enhance semantic knowledge. This approach can provide SSL with weak semantic guidance from text-image-aligned vision features [14, 19, 22, 32, 38]. Combining SSL and a vision–language foundation model is therefore a reasonable and promising approach. Image segmentation is also expected to be useful for enhancing semantic knowledge because it can provide semantically decomposed image information. However, there has been insufficient study of SSL combined with image segmentation.

In this paper, we focus on the potential of SAM-generated masks and propose a novel method for SSL, named "mask-guided attention bias". The core idea of this method is that spatial and semantic guidance created from masks is given to a ViT encoder through self-attention. Specifically, we design mask relation encoding (MRE) to encode masks from SAM to learnable attention bias as mask-guided attention bias. We add effective semantic and spatial guidance by incorporating mask-guided attention bias with MRE into SSL. Fig 1 compares general SSL and our method. Thanks to SAM's zero-shot capability, SSL with mask-guided attention bias can provide unsupervised visual representation learning.

The following summarizes the main contributions of this paper:

- Inspired by SAM's zero-shot capability, we propose mask-guided attention bias as a novel SSL method. Since our method gives a target ViT encoder spatial and semantic guidance from SAM through self-attention, it enables various SSL models to improve their feature representation.

- We design post-processing based on knowledge distillation of an SSL-trained mask-guided attention bias model to a student model that does not require SAM masks.

- Through extensive experimental evaluations, we demonstrate that our method improves the accuracy of linear probing, few-shot learning and fine-tuned evaluations on ImageNet100 [30] for general SSL models such as MAE [18] and DINO [7].

2 Related Works

Self-supervised Learning. SSL is an unsupervised representation learning method for creating a pretraining model. It is a recent breakthrough technology for mitigating supervised labeling costs for large datasets. There are two main approaches to SSL in computer vision. One is contrastive learning (CL) [6, 7, 8, 9, 10, 16, 17], which is a discriminative approach. The core idea in CL is learning "view-invariance" in multiple observations from same concepts since same concepts should produce same outputs. Specifically, the self-distillation style, which feeds multiple views to two encoders and maps one to the other by means of a predictor, has been successfully employed in many methods [1, 6, 7, 9, 16]. A key design

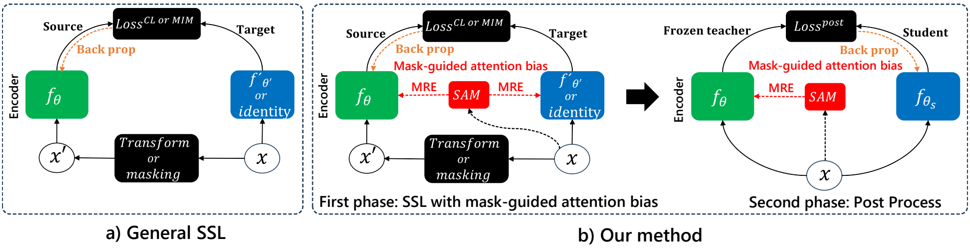


Figure 1: Architectural comparison between general SSL and our method. a) Architecture of a general SSL such as CL and MIM. b) Our method, where we incorporate mask-guided attention bias by MRE into general SSL to add effective semantic and spatial guidance from SAM.

consideration for CL is preventing collapse. For example, many methods update one of the two encoder weights with a running average of the other's weights to create a stop gradient.

The other approach is masked image modeling (MIM) [3, 18, 40, 42] as a generative method. This approach is inspired by masked language modeling in NLP [6, 11, 27, 28]. The key technique in MIM is removing masked image pixels from the input and learning to predict what was masked. ViT designs allow MIM to work effectively in terms of computation costs. MIM captures local relationships, but it is not as good as CL at capturing semantic information [26]. There are thus some MIM-based methods that incorporate a vision–language foundation model such as CLIP [29] to enhance semantic knowledge [14, 19, 22, 32, 38]. This study is inspired by these works, and to the best of our knowledge, it is the first attempt to incorporate SAM into SSL as an image segmentation foundation model.

Attention Bias. Attention bias is a technique for adding effective bias to self-attention, mainly used as relative position encoding [15, 23, 31, 39]. Image relative position encoding (iRPE) [39] considers directional relative distance modeling as well as interactions between queries and keys in self-attention. iRPE is calculated via a look-up table with learnable parameters. In the Swin Transformer V2 architecture [24], iRPE is updated to instead use a log-spaced continuous position bias. These are effective methods because vanilla self-attention cannot capture the ordering of input tokens, so they are widely used in many downstream tasks.

Attention bias is also used to add spatial conditions to self-attention. Focal attention [41] treats a bounding box as layout information and predicts the current patch token by focusing only on closely related tokens as specified by the spatial layout for image generation.

3 Method

We propose mask-guided attention bias as a novel attention-bias method for SSL. Fig 1 shows an overview of our method. Unlike general SSL, our method adds effective guidance to self-attention in the SSL process. The following subsections describe this in detail.

3.1 Preliminaries

Self-Attention for ViT. ViT is structured around self-attention [35], which can be described as mapping a query and a set of key–value pairs to an output. Given an input image token

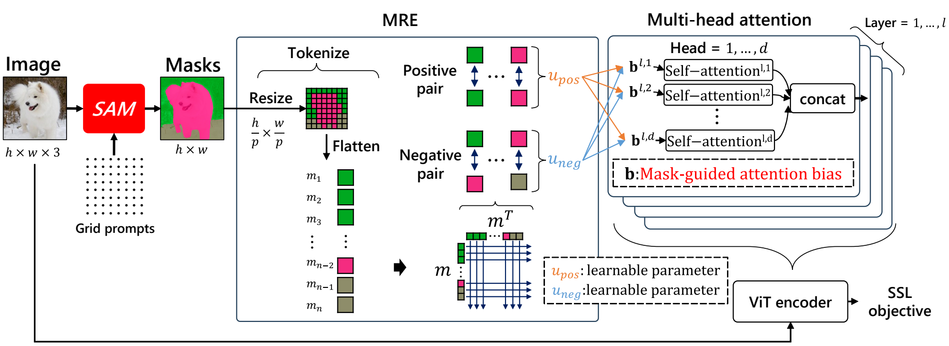


Image
h × w × 3
SAM
Grid prompts
Masks
h × w
MRE
Tokenize
Resize
$\frac{h}{p} \times \frac{w}{p}$
Flatten
m_1
m_2
m_3
m_{n-2}
m_{n-1}
m_n
Positive pair
Negative pair
u_{pos}
u_{neg}
m^T
m
u_{pos}: learnable parameter
u_{neg}: learnable parameter
Multi-head attention
Layer = 1, ..., l
Head = 1, ..., d
$\mathbf{b}^{l,1}$ Self−attention^{l,1}
$\mathbf{b}^{l,2}$ Self−attention^{l,2}
$\mathbf{b}^{l,d}$ Self−attention^{l,d}
concat
b:Mask-guided attention bias
ViT encoder
SSL objective

Figure 2: Overview of MRE on SSL with mask-guided attention bias. MRE calculates whether queries and keys belong to the same mask in self-attention and creates mask-guided attention bias accordingly.

vector $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ of n elements (where $\mathbf{x}_i \in \mathbb{R}^{d_x}$), the output $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)$ (where $\mathbf{y}_i \in \mathbb{R}^{d_y}$) and the weight coefficient α_{ij} are calculated as

$$\mathbf{y}_i = \sum_{j=1}^{n} \alpha_{ij} \cdot (\mathbf{x}_j \mathbf{W}^V), \text{ where } \alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{n} \exp(e_{ik})}. \tag{1}$$

Here, e_{ij} is a scaled dot-product attention for input $\mathbf{x}$, calculated as

$$e_{ij} = \frac{(\mathbf{x}_i \mathbf{W}^Q)(\mathbf{x}_j \mathbf{W}^K)^T}{\sqrt{d_y}}, \tag{2}$$

where $\mathbf{W}^Q, \mathbf{W}^K, \mathbf{W}^V \in \mathbb{R}^{d_x \times d_y}$ are the learnable weights. ViT adapts multi-head self-attention, an extension of self-attention in which multiple self-attention operations run. The self-attention output $\mathbf{y}$ and the input $\mathbf{x}$ are combined by the residual mechanism.

Zero-shot Segmentation by SAM. SAM is a promptable segmentation model. To get image masks in an unsupervised manner, we use SAM in automatic mask generation mode [20], which takes prompts as a regular grid of multiple points. SAM with automatic mask generation can generate reasonable masks for an image region as indicated by provided prompts. We denote SAM as S_{θ_μ} and SAM-generated masks as

$$\mathbf{M} = S_{\theta_\mu}(\mathbf{x}, \mathbf{a}), \tag{3}$$

where $\mathbf{M} \in \{x \in \mathbb{Z} \mid 0 \leq x \leq c\}^{h \times w}$, $h \times w$ is the spatial resolution, c is the maximum number of masks in an image, $\mathbf{a} \in \mathbb{R}^{2 \times z}$ is a grid prompt vector consisting of 2D coordinates, and z is the number of prompts. Note that the mask is valid where an $\mathbf{M}$ element is ≥ 1 and invalid where it is 0. In areas where masks overlap, we only leave the confident mask, so duplicate masks are not allowed in the SSL process.

3.2 SSL with Mask-guided Attention Bias

Motivation. Our motivation is exploiting masks, which are semantically and spatially decomposed information about an image for improving feature representations in SSL. Since masks are spatial information, we designed mask-guided attention bias as a mechanism to give a ViT encoder spatial guidance for self-attention. As a key insight, we focus on the relations of queries and keys in self-attention based on SAM masks, namely whether queries and keys belong to the same mask in self-attention. This relation can be derived even from zero-shot masks lacking class information. Therefore, we designed MRE to encode masks to mask-guided attention bias as a learnable attention bias. Fig 2 shows an overview of MRE.

MRE for Mask-guided Attention Bias. In preparation for MRE, we define the following relations for each pair of image tokens in self-attention: belonging to the same mask is considered *positive*, belonging to different masks is *negative*, and tokens not belonging to any mask are *unknown*. Details are presented below.

First, the mask image $\mathbf{M}$ is encoded as the mask token vector $\mathbf{m} = (m_1, \dots, m_n) \in \{x \in \mathbb{Z} \mid 0 \leq x \leq c\}^n$, which has n same elements as an image token vector by a resizing operation. Note that $n = h \times w / p^2$, where $p \times p$ is the token resolution.

Next, let mask-guided attention bias $\mathbf{b} = (b_{1,1}, \dots, b_{n,n}) \in \mathbb{R}^{n \times n}$. We then calculate the relations of each element in $\mathbf{m}$ as the relations of each mask and $\mathbf{b}$, which is shown as

$$b_{ij} = \begin{cases} u_{pos} & \text{if } (m_i = m_j) \ \& \ (m_i \neq 0, m_j \neq 0) \\ u_{neg} & \text{if } (m_i \neq m_j) \ \& \ (m_i \neq 0, m_j \neq 0) \\ 0 & \text{otherwise} \end{cases}, \tag{4}$$

where $u_{pos} \in \mathbb{R}$ is a learnable scalar for *positive* relations and $u_{neg} \in \mathbb{R}$ is a learnable scalar for *negative* relations. Note that $\mathbf{b}$ is not set as a learnable parameter for *unknown* relations. Subsequently, mask-guided attention bias is applied to scaled dot-product attention as

$$e_{ij} = \frac{(\mathbf{x}_i \mathbf{W}^Q)(\mathbf{x}_j \mathbf{W}^K)^T + b_{ij}}{\sqrt{d_y}}. \tag{5}$$

In a practical implementation, the relations of each element in $\mathbf{m}$ are calculated by the dot product of $\mathbf{m}^T$ and $\mathbf{m}$. As an advanced design, we can expand mask-guided attention bias to multiple heads and layers according to the ViT design. For instance, u_{pos} and u_{neg} have different parameters across layers and heads, like $(u_{pos}^{1,1}, \dots, u_{pos}^{l,d})$ and $(u_{neg}^{1,1}, \dots, u_{neg}^{l,d})$, where l and d are the layer number and head number, respectively. Then, we can create a multi-mask-guided attention bias $(\mathbf{b}^{1,1}, \dots, \mathbf{b}^{l,d}) \in \mathbb{R}^{n \times n \times l \times d}$. Intuitively, mask-guided attention bias for *positive* relations strengthen attention weights, and *negative* relations weaken attention weights to exploit the semantics of masks.

Application to SSL. Since mask-guided attention bias is applied to self-attention, it does not depend on the SSL method, and thus can be applied to both CL and MIM without the modification of their methods. We describe such applications below.

Let f_θ be a ViT encoder with mask-guided attention bias. For CL, we show the case of DINO [7] as a representative modern CL. Here, δ_ϕ denotes an auxiliary head connected to f_θ, with the loss function for CL with mask-guided attention bias defined as

$$L'^{cl} = L^{cl}(f_\theta \circ \delta_\phi(\mathbf{x}, \mathbf{m}), f_{\theta'} \circ \delta_{\phi'}(\mathbf{x}', \mathbf{m}')), \tag{6}$$

where L^{cl} denotes the standard DINO loss function including a regularization term to prevent collapse, $\mathbf{x}'$ and $\mathbf{m}'$ are cross-views against $\mathbf{x}$ and $\mathbf{m}$, and θ' and ϕ' are parameters of $f_{\theta'}$

and $\delta_{\phi'}$ as a teacher model obtained from the student model parameters θ and ϕ by the exponential moving average.

For MIM, we show the case of MAE [18]. With γ_ω denoting an auxiliary decoder, the loss function of MIM with mask-guided attention bias is

$$L'^{mim} = L^{mim}(f_\theta \circ \gamma_\omega(\mathbf{x}^{vis}, \mathbf{m}^{vis}), \mathbf{x}^{masked}), \tag{7}$$

where L^{mim} denotes the standard MIM loss function, the $\mathbf{x}^{vis}$ visible image token vector and the visible mask token vector $\mathbf{m}^{vis}$, and $\mathbf{x}^{masked}$ is a masked image token vector.

3.3 Knowledge Distillation as Post-processing

Necessity of Post-processing. Since applying mask-guided attention bias to a ViT encoder requires a mask like Eq. (4), a ViT encoder trained with mask-guided attention bias also needs SAM masks when it is deployed to a downstream task. This increases calculation costs according to the extent to which SAM must be applied. Therefore, we propose post-processing for SSL with mask-guided attention bias (Fig 1).

Approach. Inspired by knowledge distillation for SSL [2, 13, 25, 36], we designed post-processing to distill mask-guided attention bias trained by SSL to a student model that does not take masks. Knowledge distillation generally distills knowledge from a large teacher model to a small student model. In our method, the teacher and student models are the same model, except that the teacher model utilizes mask-guided attention bias.

Let f_{θ_s} be a post-processed ViT student model encoder that does not take mask token vectors and f_θ be a teacher model already trained with mask-guided attention bias. Note that all teacher model parameters are frozen. For DINO [7], as a representative modern CL, δ_{ϕ_s} denotes the auxiliary head of a student model. We formulate the post-processing loss as

$$L_{cl}^{post} = L_{cl}^{kd}(f_{\theta_s} \circ \delta_{\phi_s}(\mathbf{x}), f_\theta \circ \delta_\phi(\mathbf{x}, \mathbf{m})), \tag{8}$$

where L_{cl}^{kd} denotes the loss of knowledge distillation for CL [13] that removes the collapse-preventing regularization terms from L_{cl} and adapts identical-view predictions instead of cross-view predictions.

For MAE [18] as MIM, the main training objective of f_{θ_s} is matching the attention weights of f_{θ_s} and those of a teacher model f_θ already trained with mask-guided attention bias. We formulate the attention loss as

$$L_{att} = \mathrm{MSE}(\mathbf{A}^{\mathrm{s}}, \mathbf{A}^{\mathrm{t}}), \tag{9}$$

where $\mathbf{A}^s = \{\alpha^s_{i=1,j=1}, \dots, \alpha^s_{i=n,j=n}\}^{n\times n}$ and $\mathbf{A}^t = \{\alpha^t_{i=1,j=1}, \dots, \alpha^t_{i=n,j=n}\}^{n\times n}$ denote the attention weights of the corresponding student and teacher layers, and MSE is a mean squared error function. Because mask-guided attention bias directly affects an attention mechanism, attention loss effectively distills teacher knowledge. Let γ_{ω_s} be the auxiliary decoder of a student model, we formulate the post-processing loss as

$$L_{mim}^{post} = L_{mim}(f_{\theta_s} \circ \gamma_{\omega_s}(\mathbf{x}^{vis}), \mathbf{x}^{masked}) + \lambda L_{att}. \tag{10}$$

Following Ref. [36], we distill only teacher knowledge through attention loss. $\lambda > 0$ controls the attention loss weight.

4 Experiments

In experiments, we performed SSL with mask-guided attention bias and ViT encoder post-processing. Then, we performed to linear probing evaluation to measure the decodability of its feature representations. To evaluate label efficiency, we performed 1% and 10% labeled data fine-tuned evaluations for few-shot image recognition. As a practical downstream task evaluation, we performed full-data fine-tuning. In addition, we conducted various ablation studies and analyses for deeper understanding.

Experimental Setup. We selected DINO [7] and MAE [18] as baseline CL and MIM methods, respectively. We mainly used ViT-B and ViT-S encoders, and performed experiments on ImageNet100 [30]. ImageNet100 is a subset of ImageNet1k, containing 100 classes. There are about 1,300 samples for each class. We used publicly available ViT/H SAM trained with SA-1B dataset [20]. As key hyperparameters for our method, we set the maximum number of masks $c = 5$, the number of prompts $z = 1024 (32 \times 32)$, and $\lambda = 1$ in post-processing. More details are provided in the supplementary material.

4.1 Image Classification

First, we compared the baseline SSL and our method on ImageNet100. Table 1 shows the results for ImageNet100. Adding mask-guided attention bias resulted in accuracy gains over the use of vanilla MAE and DINO. Our method resulted in particularly large accuracy gains for linear probing in MAE. This shows that our method can provide a ViT encoder semantic guidance that improves the SSL process, especially for MAE, which has low representation abstraction. Our method also improves fine-tuning accuracy for both MAE and DINO in almost all of architectures, token resolutions and training epochs.

Method	Arch	Epochs	Linear	Fine-tuned on		
				1%	10%	100%
DINO	ViT-S/16	300	81.9	68.6	82.3	88.0
DINO + our method	ViT-S/16	300	81.7 (↓ 0.2)	68.7 (↑ 0.1)	82.8 (↑ 0.5)	88.7 (↑ 0.7)
DINO	ViT-S/16	600	82.5	70.9	82.8	88.3
DINO + our method	ViT-S/16	600	81.3 (↓ 1.2)	71.2 (↑ 0.3)	83.0 (↑ 0.2)	88.3
DINO	ViT-S/8	300	84.9	70.5	83.6	89.1
DINO + our method	ViT-S/8	300	84.2 (↓ 0.7)	69.2 (↓ 1.3)	84.1 (↑ 0.5)	89.5 (↑ 0.4)
MAE	ViT-B/16	1600	73.3	64.8	78.5	88.1
MAE + our method	ViT-B/16	1600	77.1 (↑ 3.8)	65.0 (↑ 0.2)	79.0 (↑ 0.5)	88.6 (↑ 0.5)
MAE	ViT-B/16	3200	76.9	66.2	79.8	88.4
MAE + our method	ViT-B/16	3200	79.2 (↑ 2.3)	66.9 (↑ 0.7)	80.2 (↑ 0.3)	88.7 (↑ 0.3)
MAE	ViT-B/8	1600	78.1	76.3	85.7	89.9
MAE + our method	ViT-B/8	1600	81.3 (↑ 3.2)	76.6(↑ 0.3)	85.8 (↑ 0.1)	90.2 (↑ 0.3)

Table 1: Evaluation results for ImageNet100

SSL method	Mask-guided attention bias	Post-processing	Linear
MAE	-	-	73.3
MAE + our method (1st phase)	✓	-	77.2 (↑ 3.9)
MAE + our method	-	-	76.0 (↑ 2.7)
MAE + our method (2nd phase)	-	✓	77.1 (↑ 3.8)

Table 2: Ablation study for post-processing

Prompt setting	Average mask ratio (%)	SSL performance (linear prob.)
8 x 8 grid	70.4	76.6
16 x 16 grid	72.0	77.0
32 x 32 grid	72.7	77.1

Table 3: Comparison with SAM prompt settings

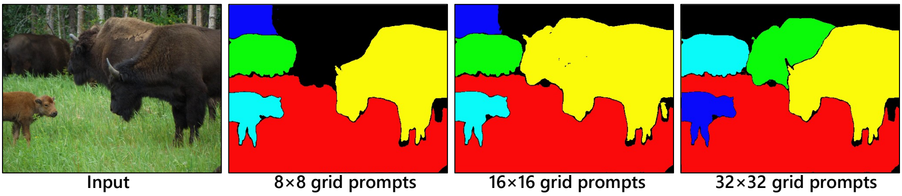


Figure 3: Examples of SAM masks with different prompts

SSL method	*Positive* relations	*Negative* relations	Linear
MAE	-	-	73.3
MAE + our method	✓	-	76.8 (↑ 3.5)
MAE + our method	-	✓	74.1 (↑ 0.8)
MAE + our method	✓	✓	77.1 (↑ 3.8)

Table 4: Ablation study for mask relation encoding design

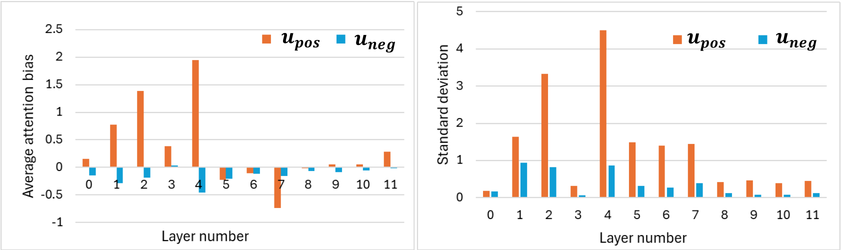
4.2 Ablation Study

We conducted various ablation studies to investigate the effectiveness of our method. Unless otherwise noted, we used MAE with ViT-B/16 trained by 1600 epochs in these ablation studies.

Importance of Post-processing. We evaluated the importance of our post-processing method in linear probing. Table 2 compares the our SSL method with and without post-processing, showing that the post-processing maintains first-phase performance as in the case without mask-guided attention bias. This means our post-processing can distill guidance as mask-guided attention bias to a student model. Surprisingly, a pretrained model without our post-processing still outperforms the baseline, indicating that our guidance encourages not just self-attention, but also the whole SSL process.

Evaluation of Prompt Settings Given to SAM. SAM prompts are important parameters in our method because the mask quality depends on the prompts. Therefore, we compared several SAM prompts and evaluated SSL performance. Table 3 shows the results and Fig 3 shows the examples of SAM masks. Note that SAM outputs more detailed masks as the number of points increases, providing more effective guidance to SSL.

Mask Relation Encoding Design. We compared the effectiveness of mask-guided attention bias for *positive* and *negative* relations. Table 4 shows the results of linear probing for each attention bias. We note that the attention bias for *positive* relations is more effective for SSL. However, the attention bias for *negative* relations too contributes to SSL, so our MRE design exploits both *positive* and *negative* relations.



a) Average value of attention bias b) Standard deviation of attention bias

Figure 4: Analysis of trained mask-guided attention bias in MAE with ViT-B/16. a) Averaged value across each head of mask-guided attention bias. b) Standard deviation across each head of mask-guided attention bias.

SSL method	Epochs	multi heads setting	multi layers setting	Linear
MAE	1600	-	-	73.3
MAE + fixed attention bias	1600	-	-	72.8 (↓ 0.5)
MAE + learnable attention bias	1600	-	-	74.1 (↑ 0.8)
MAE + learnable attention bias	1600	✓	-	74.6 (↑ 1.3)
MAE + learnable attention bias	1600	✓	✓	77.1 (↑ 3.8)

Table 5: Comparison with mask-guided attention bias designs. Note that 1 is set for *positive* relations and -1 is set for *negative* relations as "fixed attention bias.

4.3 Further Analysis

Trained Mask-guided Attention Bias. We observed mask-guided attention bias trained by MAE. Fig 4 shows the trained mask-guided attention bias averaged across the heads. Generally, attention biases for positive relations are larger than those for negative relations, and negative relations are negative values. This indicates that self-attention emphasizes attention with respect to tokens belonging to the same mask, and that mask-guided attention bias can provide semantic guidance based on those masks. In addition, trained mask-guided attention bias shows diversity across the heads, seemingly contributing to multi-head self-attention. As an interesting phenomenon, we observe that trained mask-guided attention biases have larger shallow layers than deep layers. This suggests that the ViT encoder aggressively exploits semantic guidance from SAM when the image's original spatial information is more preserved. Based of this observation, we evaluate the design of mask-guided attention bias. Table 5 shows that the learnable setting and an expanding design to multiple heads and layers effectively contribute the SSL performance.

5 Conclusion

Focusing on the potential of SAM-generated masks, we proposed mask-guided attention bias as a novel SSL method. We showed that SSL with mask-guided attention bias can improve feature representations over those of vanilla DINO and MAE. Extensive ablation studies showed that mask-guided attention bias encourages SSL by exploiting semantics from masks and effectively works at each head and layer.

Limitations: Due to a limited computation budget, we performed experiments on small datasets such as ImageNet100 and limited ViT architectures. In the future, we will evaluate our method on larger datasets and more varied ViT architectures, such as the Swin transformer [23].

References

[1] Mahmoud Assran, Mathilde Caron, Ishan Misra, Piotr Bojanowski, Florian Bordes, Pascal Vincent, Armand Joulin, Michael Rabbat, and Nicolas Ballas. Masked siamese networks for label-efficient learning. In *ECCV*, 2022.

[2] Yutong Bai, Zeyu Wang, Junfei Xiao, Chen Wei, Huiyu Wang, Alan L. Yuille, Yuyin Zhou, and Cihang Xie. Masked autoencoders enable efficient knowledge distillers. In *CVPR*, 2023.

[3] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. BEit: BERT pre-training of image transformers. In *ICLR*, 2022.

[4] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, S. Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie S. Chen, Kathleen A. Creel, Jared Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren E. Gillespie, Karan Goel, Noah D. Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas F. Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, O. Khattab, Pang Wei Koh, Mark S. Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir P. Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Benjamin Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, J. F. Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Robert Reich, Hongyu Ren, Frieda Rong, Yusuf H. Roohani, Camilo Ruiz, Jack Ryan, Christopher R'e, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishna Parasuram Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei A. Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the opportunities and risks of foundation models. *arXiv:2108.07258*, 2021.

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *NeurIPS*, 2020.

[6] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *NeurIPS*, 2020.

[7] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021.

[8] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020.

[9] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *CVPR*, 2021.

[10] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *ICCV*, 2021.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019.

[12] 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 *ICLR*, 2021.

[13] Quentin Duval, Ishan Misra, and Nicolas Ballas. A simple recipe for competitive low-compute self supervised vision models. *arXiv:2301.09451*, 2023.

[14] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. In *CVPR*, 2023.

[15] Benjamin Graham, Alaaeldin El-Nouby, Hugo Touvron, Pierre Stock, Armand Joulin, Herve Jegou, and Matthijs Douze. Levit: A vision transformer in convnet's clothing for faster inference. In *ICCV*, 2021.

[16] 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 *NeurIPS*, 2020.

[17] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020.

[18] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022.

[19] Zejiang Hou, Fei Sun, Yen-Kuang Chen, Yuan Xie, and Sun-Yuan Kung. Milan: Masked image pretraining on language assisted representation. *arXiv:2208.06049*, 2022.

[20] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment anything. In *ICCV*, 2023.

[21] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[22] Yanghao Li, Haoqi Fan, Ronghang Hu, Christoph Feichtenhofer, and Kaiming He. Scaling language-image pre-training via masking. In *CVPR*, 2023.

[23] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021.

[24] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, and Baining Guo. Swin transformer v2: Scaling up capacity and resolution. In *CVPR*, 2022.

[25] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *arXiv:2304.07193*, 2023.

[26] Namuk Park, Wonjae Kim, Byeongho Heo, Taekyung Kim, and Sangdoo Yun. What do self-supervised vision transformers learn? In *ICLR*, 2023.

[27] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pretraining. 2018.

[28] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, DarioD Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

[29] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya 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.

[30] 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. *IJCV*, 115 (3):211–252, 2015.

[31] Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. In *ACL*, 2018.

[32] Mannat Singh, Quentin Duval, Kalyan Vasudev Alwala, Haoqi Fan, Vaibhav Aggarwal, Aaron Adcock, Armand Joulin, Piotr Dollar, Christoph Feichtenhofer, Ross Girshick, Rohit Girdhar, and Ishan Misra. The effectiveness of mae pre-pretraining for billion-scale pretraining. In *ICCV*, 2023.

[33] Sebastian Thrun. Is learning the n-th thing any easier than learning the first? In *NeurIPS*, 1995.

[34] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. Training data-efficient image transformers & distillation through attention. In *ICML*, 2021.

[35] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017.

[36] Shaoru Wang, Jin Gao, Zeming Li, Xiaoqin Zhang, and Weiming Hu. A closer look at self-supervised lightweight vision transformers. In *ICML*, 2023.

[37] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *ICCV*, 2021.

[38] Longhui Wei, Lingxi Xie, Wengang Zhou, Houqiang Li, and Qi Tian. Mvp: Multimodality-guided visual pre-training. In *ECCV*, 2022.

[39] Kan Wu, Houwen Peng, Minghao Chen, Jianlong Fu, and Hongyang Chao. Rethinking and improving relative position encoding for vision transformer. In *ICCV*, 2021.

[40] 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 *CVPR*, 2022.

[41] Zuopeng Yang, Daqing Liu, Chaoyue Wang, Jie Yang, and Dacheng Tao. Modeling image composition for complex scene generation. In *CVPR*, 2022.

[42] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. In *ICLR*, 2022.