DEEP DYNAMIC AUTOENCODER FOR VISION BERT PRETRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

Recently, masked image modeling (MIM) has demonstrated promising prospects in self-supervised representation learning. However, existing MIM frameworks recover all masked patches equivalently, ignoring that the reconstruction difficulty of different patches can vary sharply due to their diverse distance from visible patches. In this paper, we propose Deep Dynamic AutoEncoder (DDAE), a novel MIM framework that dynamically focuses on patch reconstructions with different degrees of difficulty at different pretraining phases and depths of the model. In addition to raw pixel regression, DDAE performs dynamic feature self-distillation for intermediate layers to learn semantic information. Our methodology provides more locality inductive bias for ViTs, especially in deep layers, which inherently makes up for the absence of local prior for self-attention mechanism. Moreover, our core design deep dynamic supervision can be migrated into existing MIM methods (*e.g.*, MAE, BEiT-v2) seamlessly. The Experimental results demonstrate the effectiveness of our approach. As a tokenizer-free framework, the base-size DDAE can achieve 83.5% top-1 accuracy with only 100 epochs pretraining, surpassing MAE and BEiT pretrained for 800 epochs. For a longer pretraining schedule, DDAE achieves 84.3% top-1 accuracy on Imagenet-1K, and 49.3% mIoU on ADE20K for semantic segmentation.

1 INTRODUCTION

Aided by the rapid gains in hardware, deep learning has ushered in the era of big models and big data. Along with the ever-growing model capacity, the demand for data can easily reach hundreds of millions (Dosovitskiy et al., 2020), which is not publicly accessible for labeled data. Self-Supervised Learning(SSL) frameworks, such as DINO (Caron et al., 2021), MOCO (Chen et al., 2021), BeiT (Bao et al., 2021), etc., have grown in concern in vision model pretraining without the need for labels. In particular, the recently proposed Masked Image Modeling(MIM) methods (He et al., 2022; Xie et al., 2022b; Dong et al., 2021; Wei et al., 2022b; Chen et al., 2022b; Dong et al., 2022; Chen et al., 2022c) have shown remarkably impressive performance in a variety of vision tasks, demonstrating the promise to unify computer vision and natural language processing(NLP) pretraining (Peng et al., 2022a).

Inspired by BERT (Devlin et al., 2018) in NLP, MIM pretrains the encoder by reconstructing the masked image patches from visible patches. Existing MIM methods can be divided into two categories according to the need for an additional tokenizer. **Two-stage methods:** Represented by the pioneer work BEiT (Bao et al., 2021), two-stage methods firstly transform image patches into semantic visual tokens through a pre-trained tokenizer, then reconstruct the corresponding tokens of masked image patches to pretrain the encoder. Tokenizer needs to be offline pretrained with fixed model architectures and extra data (Zhang et al., 2019; Ramesh et al., 2021; Radford et al., 2021), some methods further require an off-the-shelf DNN as teacher to distill tokenizer pre-training (Peng et al., 2022a). **One-stage methods:** Taking the recent work MAE (He et al., 2022) as a representative, MAE constructs an asymmetric encoder-decoder structure, and directly performs raw pixel regression of masked image patches. One-stage methods are tokenizer-free, the reconstruction target includes raw pixels, features (*e.g.*, HOG) and self-distillation targets (He et al., 2022; Xie et al., 2022b; Wei et al., 2022a; Chen et al., 2022c). MIM methods enable Vision Transformers(ViTs)to learn rich visual representations and show great potential in various downstream tasks.

Naturally, we ask why MIM works. Facilitated by self-attention mechanism, ViTs excel at modeling long-range dependencies, but different from CNN, ViTs' lack of local prior(*i.e.*, a pixel is more related to its neighbors than the distant pixels) largely leads to its need for large dataset pretraining (*e.g.*, JFT-300M) or special training techniques (*e.g.*, a DeiT-style distillation method (Touvron et al., 2021b)). Therefore, we conjecture that the success of MIM comes from the fine-grained reconstruction of masked image patches. The generation task built by MIM enables ViTs to pay more attention to vicinity besides using global semantic reasoning. Thus, it inherently makes up for the absence of local prior in the structure of ViTs, which is essentially different from the discriminative task in supervised pretraining or contrastive learning. As analyzed in Xie et al. (2022a), compared with traditional supervised pretraining and contrastive learning, MIM makes the average attention distance of ViTs closer, especially in deeper blocks. This observation partially supports our conjecture that it is beneficial to inject local priors into ViT through pretraining.

However, success comes with remaining obstacles. At present, all MIM frameworks recover all patches equivalently, ignoring the fact that the reconstruction difficulty of different patches can vary sharply, and the semantic reasoning and local perception ability required for patches are not the same. Intuitively, recovering patches with more visible patches around will be simpler, as long as the model has sufficient local perception ability. In contrast, reconstruction with few visible patches around requires the model to have strong semantic reasoning ability, given that there is little access to neighboring information. Treating all pixels equally will neglect this demand for different properties of the model, inevitably limiting its representation ability. On the other hand, layers at different depths and training phases naturally learn features at different levels. Therefore, we ask if there is a way to focus on objectives with diverse characteristics as training progresses so that better representations can be learned overall.

Motivated by this observation and answering the question above, we propose a simple yet effective SSL framework named **D**eep **D**ynamic **A**uto**E**ncoder (DDAE). With deep dynamic supervision, the model dynamically focuses on different patches at different depths and training phases. First, it attaches importance to patches close to visible ones, which only demands fair local perception to recover. Then compared with the beginning of training, the model further pays attention to the distant and difficult patch aided by the gradually reconstructed vicinity. Thus, from near to distant, dynamic targets are introduced for layers with distinct characteristics, leading to a better representation of the overall structure. Specifically, we first define the reconstruction difficulty according to the distance between masked patches and visible ones, generating a distance map. Then, we exert different supervision signals (which are controlled by learnable parameters β) for intermediate layers. As the training progresses, the model dynamically focuses on different regions with the update of β. Note that β of layers are independent of each other. Although most traditional deep supervision methods build consistent targets to alleviate the vanish of gradients and accelerate convergence, this paper shows another potential use of deep supervision to make the network more diverse.

Since the pixel representations of perceptually similar images might be very different, only pixel-level reconstruction can be vulnerable to the perceptual difference between images (Dong et al., 2021). Besides pixel regression, we further apply feature self-distillation to provide semantic information. We directly align the encoder's intermediate features with the corresponding features of the momentum encoder in deep dynamic supervision, where the momentum encoder is updated by Exponential Moving Average (EMA) (Grill et al., 2020; He et al., 2020). It is worth noting that the feature distillation target here is the corresponding feature of the momentum encoder, while the regression targets for intermediate layers are all raw pixels. Dynamic Loss is applied for both feature self-distillation and pixel regression.

By analyzing the characteristics of the pretrained encoder, we show that DDAE can effectively strengthen ViTs' attention to local patterns, establishing a pretraining task that benefits ViTs from structural priority. DDAE forms target perception attention effectively in the very early stage, while MAE still behaves to be rudimentary (Fig 3).

Furthermore, we directly migrate the core design deep dynamic supervision to the representative methods in one-stage and two-stage, MAE (He et al., 2022) and BEiT-v2 (Peng et al., 2022a) respectively, surpassing original methods with nontrivial margins. Since deep dynamic supervision does not introduce any additional structure, it can be used as a general MIM pretraining plug-and-play module.

The contributions are summarized as follows:

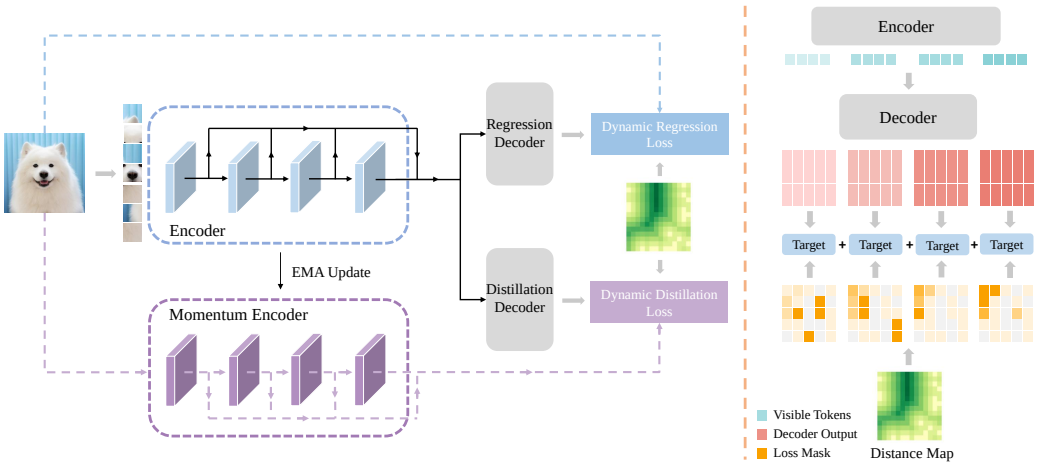
- We propose DDAE, a one-stage tokenizer-free framework for self-supervised representation learning. Our core design deep dynamic supervision can be migrated to existing MIM approaches seamlessly, providing more local priors for ViTs to make up for their inherent structural defects.
- We utilize deep supervision to guide the diversification of different components of the network through inconsistent signals. Although deep supervision does not perform well in modern supervised learning, this paper makes a small step towards unleashing the potential of deep supervision in MIM.
- The experimental results demonstrate the effectiveness of the proposed DDAE, we achieve state-of-the-art performance among tokenizer-free methods and have competitive performance with many two-stage methods.

2 Related Work

Local Prior for ViTs. Aided by the global receptive field of multi-head self-attentions(MSAs) mechanism, ViTs (Dosovitskiy et al., 2020) have significantly advanced computer vision tasks such as image classification (Wu et al., 2020; Touvron et al., 2021a; Yuan et al., 2021), object detection (Carion et al., 2020; Zhu et al., 2020; Beal et al., 2020), and semantic segmentation (Zheng et al., 2021).However, since self-attention does not have CNN's local prior, it often ignores local feature details, which largely leads to ViT's demand for large dataset pretraining. To solve, DeiT (Touvron et al., 2021a) transfers CNN-based knowledge to vision transformer through a distillation token. T2T-ViT (Yuan et al., 2021) proposed using a tokenization module to recursively reorganize the image to consider neighboring pixels. Local MSAs (Yang et al., 2019; Liu et al., 2021; Chu et al., 2021) are proposed to calculate self-attention only within small windows, achieving better performance than global MSAs not only on small datasets but also on large datasets, *e.g.*, ImageNet-21K. Since local prior is the natural inductive bias of convolution, there are also works aimed at integrating convolution and vision transformer. Works like PVT (Wang et al., 2021)and CvT (Wu et al., 2021) insert spatial reduction or convolution before global self-attention, yielding the merits of self-attention and convolution. MixFormer (Chen et al., 2022a) and Conformer (Peng et al., 2021) integrate the advantages by building parallel interactions. It should be noted that all above methods fundamentally change ViTs' basic structure, bringing negative effects such as restricted receptive field or increased computational cost of downstream tasks. Different from them, our work focuses on providing local prior for vanilla ViTs without modifying the structure. By dynamically changing the pretraining mechanism of MIM, we enable ViTs to pay more attention to local vicinity without introducing any extra structural modification.

Masked Image Modeling. Inspired by BERT (Devlin et al., 2018) in NLP, MIM learn representation by reconstructing masked image patches from visible patches. Existing MIM methods can be divided into two categories according to the need for the additional tokenizer. **Two-stage methods:** Represented by the pioneering work BEiT (Bao et al., 2021), two-stage methods (Bao et al., 2021; Dong et al., 2021; Wei et al., 2022b; Chen et al., 2022b; Peng et al., 2022b) firstly transform image patches into semantic visual tokens through a pretrained discrete VAE (Ramesh et al., 2021)as visual tokenizer, then the corresponding tokens of masked image patches are reconstructed to pretrain the encoder. Nevertheless, the tokenizer needs to be offline pretrained with fixed model architectures and extra data (Zhang et al., 2019; Ramesh et al., 2021; Radford et al., 2021), some methods even further require an off-the-shelf DNN as teacher to distill tokenizer pre-training (Peng et al., 2022b), which complicates the process and increases cost. **One-stage methods:** MAE (He et al., 2022)constructs an asymmetric encoder-decoder structure, and directly performs raw pixel regression of masked image patches. SimMIM (Xie et al., 2022b) allows hierarchical transformers such as Swin (Liu et al., 2021) to be directly applied to MIM. MaskFeat (Wei et al., 2022a), Data2vec (Baevski et al., 2022), BootMAE (Dong et al., 2022), sDAE (Chen et al., 2022c) explored the choice of reconstruction targets. Our framework DDAE belongs to one-stage methods, which are tokenizer-free.

Deep Supervision. Deep supervision methods are proposed to accelerate model's convergence and alleviate the problem of gradient vanishment. Through applying aux layers to transmit the super-



(a) Overall Framework of DDAE

(b) Deep Dynamic Supervision

Figure 1: Overview of DDAE pre-training. (a) the overall framework; (b) deep dynamic supervision. Lighter color indicates that the signal comes from the shallower layer of the encoder. Targets include raw pixel regression and feature self-distillation. Best viewed in color.

vision signal to the shallow layers, deep supervision has been used in early classification models (Szegedy et al., 2015; Lee et al., 2015) and extended to other visual recognition tasks (Xie & Tu, 2015; Newell et al., 2016; Zhang et al., 2018b; Mosinska et al., 2018; Zhang et al., 2018a). Despite these advances, modern CNN classification models rarely use auxiliary classifiers since directly appending simple auxiliary classifiers on top of the early layers of a modern network hurts its performance (Huang et al., 2017). Zhang et al. (2022) proposed Contrastive Deep Supervision to use contrastive learning signals for intermediate layers, but the signals of intermediate layers are still the same. Our approach introduces dynamic supervision for the intermediate layer in self-supervised pretraining. Different from them, we introduce inconsistency in the intermediate signals for the first time. To the best of our knowledge, this paper is also the first to apply deep supervision to the MIM task. State-of-the-art results show the great potential of deep supervision in the application of MIM.

3 METHODS

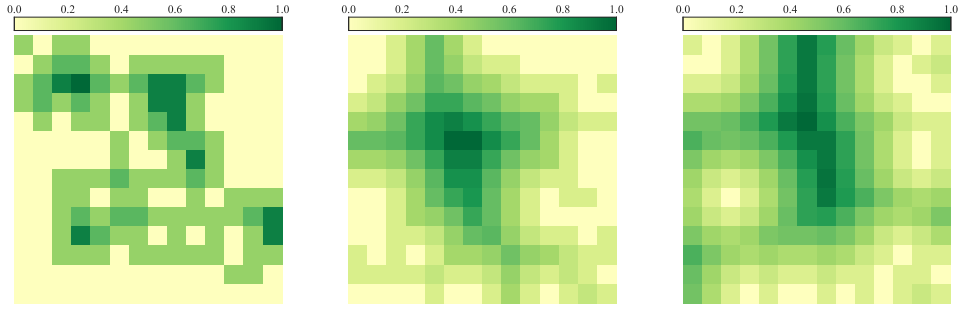
In this section, we firstly elaborate on the basic framework of MIM. Then we introduce our proposed DDAE's two core designs termed Deep Dynamic Supervision and Deep Self-Distillation in Section 3.2 and Section 3.3 respectively. Our framework is illustrated in Fig 1.

3.1 PRELIMINARY

Formally, MIM firstly divides the input image $X \in \mathbb{R}^{H\times W\times C}$ into non-overlapping flattened patches $\mathbf{x} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N]$, where $\mathbf{x}_i \in \mathbb{R}^{P^2C}$ according to the patch size P, and $N = (H \times W)/P^2$ is the number of patches. It then samples a random binary mask $\mathbf{m} = [m_1, m_2, ..., m_N]$, where $m_i \in \{0, 1\}$ to mask a portion of the flattened patches. The masked patches $\mathbf{x}_M \triangleq \mathbf{x} \odot \mathbf{m}$ are discarded (He et al., 2022) or substituted by a learnable token `[MASK]` (Bao et al., 2021), and the rest patches $\mathbf{x}_{1-M} \triangleq \mathbf{x} \odot (1 - \mathbf{m})$ are used to reconstruct the dropped features or images to learn rich representations. The optimization target of MIM can be formulated as follow:

$$\min_{\theta,\phi} \mathop{\mathbb{E}}_{X\sim D} \mathcal{M}(d_\phi([f_\theta(\mathbf{x}_{1-M}), [\text{MASK}] \odot \mathbf{m}]), \mathbf{x}_M) \tag{1}$$

where $\odot$ denotes element-wise multiplication; $f_\theta(\cdot)$ and $d_\phi(\cdot)$ are encoder and decoder respectively; $\mathcal{M}(\cdot,\cdot)$ is the similarity measurement, which varies in different works, *e.g.*, l_2-distance in pixel space (He et al., 2022), perceptual loss in codebook space (Dong et al., 2021) or self-distillation loss in feature space (Chen et al., 2022c). In our work, we use the l_2-distance as our measurement, $\mathcal{M}(a, b) = ||a - b||^2$, for both of the pixel reconstruction and self-distillation. To simplify the formulation, we ignore the mask token term and use $d_\phi(\cdot)$ to represent $d_\phi([\cdot, [\text{MASK}] \odot \mathbf{m}])$.



(a) 50% (b) 75% (c) 90%

Figure 2: Distance map (normalized to [0, 1]) under different mask ratio. Darker color means larger distance from visible patches. High mask ratio leads to diverse distance for masked patches.

3.2 Deep Dynamic Supervision

Dynamic. Firstly, we define the difficulty of reconstruction according to the distance between each masked patch and visible patches, generating a *distance map* with a distance transform $D(\cdot)$. For each masked token ($m_i = 1$), the distance transform assigns a number that is the Euclidean distance between that token and the nearest unmasked token in 2D space. Naturally, it is difficult to recover a patch that is far from visible ones and demands stronger semantic reasoning ability. On the contrary, reconstruction of a patch with visible ones nearby only requires fair local perception.

As shown in Fig 2, considering that existing MIM methods often use a high mask ratio like 75% (which is proved to be critical to MIM's success (He et al., 2022)), values of the distance map vary diversely. Since the distance map only depends on the input mask, it is data-specific and cost-free.

To guide the model to focus on different patches (corresponding to the requirement of distinctive properties) in different training phases, we propose $h_\beta(\cdot)$ to learn a dynamic coefficient β of the distance map to generate loss weight, directly applying dynamic focus to the loss weight of corresponding patches. h_β is derived as follows:

$$h(m_i|\beta) = \begin{cases} \dfrac{\exp(D(m_i)^\beta)}{\sum_{m_j=1}\exp(D(m_j)^\beta)} & , m_i = 1 \\ 0 & , m_i = 0 \end{cases} \tag{2}$$

We re-scale the loss weight map into $[0, 1]$ according to the max value of the weights. When $\beta > 0$, larger distance leads to greater loss weight, so the model pays more attention to the reconstruction of difficult patches. Increasing the value of β can exacerbate this trend. Conversely, when $\beta < 0$, larger distance leads to smaller loss weight, and decreasing β results in more importance attached to simple patches. Along with the changing of β, the model dynamically focuses on patches with diverse degrees of recovery difficulty.

Deep supervision. As commented in the introduction, the reconstruction of patches with diverse difficulties requires distinct characteristics of the model, and layers at different depths naturally learn features at different levels. Therefore, we apply dynamic loss to the intermediate layers. Taking the standard ViT with B blocks as the encoder, we divide the blocks into K groups and extract the features at the end of each group. For example, we set $K = 4$, then we extract output features of block 3, 6, 9 and 12 and feed them into the decoder respectively to recover masked patches. K is also called *supervisory number*, since K is equal to the number of supervisory signals. Then the loss function of pixel reconstruction L_{pixel} is derived as follows:

$$L_{pixel}(\theta, \phi, \beta) = \sum_{i=1}^{K} \lambda_i h_{\beta_i}(\mathbf{m}) \odot ||d_\phi(f_\theta^{(g_i)}(\mathbf{x}_M)) - \mathbf{x}_{1-M}||^2 \tag{3}$$

where $f^{(i)}(\cdot)$ denotes to the output features of block-i in encoder, and $g_i = \frac{B}{K}i$ denotes to the group index; $\boldsymbol{\lambda}$ are the scale factors of each group, and set to $[0.75, 0.75, 0.75, 1]$ by default. It should be emphasized that β of different layers are independent of each other.

Deep dynamic supervision does not introduce any additional structure, so it can be seamlessly migrated to existing MIM structures. We discussed its universality in Section 4.2 where we migrated it to the representative MIM methods of one-stage and multi-stage.

3.3 DEEP SELF-DISTILLATION

In addition to raw pixel regression, the layer-by-layer correspondence between intermediate features is more suitable for the design of deep dynamic supervision. Therefore, we designed Deep Self-Distillation based on BootMAE (Dong et al., 2022) to further strengthen the model through high-level semantic features. Specifically, the momentum encoder provides the features of masked patches of each layer as the target of self-distillation, so that the intermediate features of each layer corresponding to the encoder have their own targets of self-distillation. The momentum encoder is updated by the encoder using the exponential moving average method (EMA). Deep dynamic supervision is shared with raw pixel regression. Note that the regression of raw pixels is one-vs-all, the features of all layers will be reconstructed by one regression decoder. The feature self-distillation is all-vs-all, which means the shallow features of the encoder are self-distilled by the corresponding shallow features of the momentum encoder, and the deep features are self-distilled by the corresponding deep features. Pixel regression and feature self-distillation use separate decoders, which both consist of two-layer transformer blocks. Since the decoders are light-weighted, the additional cost brought by deep supervision is acceptable. Now the deep self-distillation target formula is:

$$L_{distill}(\theta, \beta) = \sum_{i=1}^{K} \lambda_i h_{\beta_i}(\mathbf{m}) \odot ||f_{\theta}^{(g_i)}(\mathbf{x}_M) - f_{\theta'}^{(g_i)}(\mathbf{x}_M)||^2 \tag{4}$$

where $f_\theta(\cdot)$ and $f_{\theta'}(\cdot)$ are encoder and momentum encoder respectively. $\boldsymbol{\lambda}$ are the scale factors of each group, and set to $[0.75, 0.75, 0.75, 1]$ by default.

We add an L_2 regularization term to constrain the magnitude of β, and the overall loss function is:

$$L = L_{pixel} + L_{distill} + \lambda \sum_i ||h_{\beta_i}(\mathbf{m})|| \tag{5}$$

where λ is the scale factor to tune the L_2 regularization term and is set to 0.5 by default.

4 EXPERIMENTS

4.1 IMPLEMENTATION

We conduct experiments on ImageNet-1K without labels as the pretraining data for self-supervised learning. The input resolution is set as 224×224 during pretraining and partitioned into 16×16 size patches. We pretrain the standard ViT small and base architectures, i.e., ViT-S/16 and ViT-B/16. The pixel regression decoder and feature distillation decoder both consist of 2 transformer blocks, along with an extra linear projection head for predictions. The dimension for the pixel regression decoder is 512 while for feature distillation decoder is set the same as the encoder. We use block-wise masking with a ratio of 75%. The data augmentation is only standard random cropping and horizontal flipping. All β are initialized as -0.5 by default. The pretraining settings are almost the same as BootMAE (Dong et al., 2022) (See Appendix A for details).

4.2 IMAGE CLASSIFICATION

We evaluate both fine-tuning accuracy and linear probing accuracy on ImageNet-1k. Table 1 presents the comparison with previous state-of-the-art MIM-based methods including one-stage and two-stage. As a tokenizer-free framework, our DDAE can not only achieve higher performance, but also accelerate the convergence speed of pretraining. In particular, with only 100 epochs pre-training, DDAE can achieve comparable performance with MAE using 1600 epochs pre-training and surpass 800 epochs pre-trained BEiT. Furthermore, with a longer pretraining schedule, DDAE achieves 84.3% top-1 accuracy, developing a new state-of-the-art on ImageNet-1K among one-stage self-supervised methods. Note that despite SdAE reaching comparable results with DDAE when pre-trained for 100 epochs, our method outperforms it with + 0.3% gains with 800 epoch schedule, demonstrating our merits of higher upper limit besides rapid convergence.

Table 1: Image classification accuracy (%) comparison on ImageNet-1K of different methods with ViT-B as backbone. We report the fine-tuning and linear probing accuracy and our method DDAE outperforms previous self-supervised methods.

Methods	Extra Tokenizer	Pretraining Epochs	Finetune	Linear
DINO (Caron et al., 2021)	contrastive	300	82.8	78.2
MoCo-v3 (Chen et al., 2021)	contrastive	300	83.2	76.7
SplitMask (El-Nouby et al., 2021)	w/	300	83.6	N/A
CAE (Chen et al., 2022b)	w/	300	83.6	64.1
MaskFeat (Wei et al., 2022a)	w/o	300	83.6	N/A
BootMAE (Dong et al., 2022)	w/o	100	83.0	N/A
SdAE (Chen et al., 2022c)	w/o	100	83.5	60.3
DDAE(ours)	w/o	100	**83.5**	60.3
BEiT (Bao et al., 2021)	w/	800	83.2	56.7
MAE (He et al., 2022)	w/o	800	83.4	64.4
MAE (He et al., 2022)	w/o	1600	83.6	68.0
SimMIM (Xie et al., 2022b)	w/o	800	83.8	56.7
CAE (Chen et al., 2022b)	w/	1600	83.9	70.4
MaskFeat (Wei et al., 2022a)	w/o	1600	84.0	N/A
BootMAE (Dong et al., 2022)	w/o	800	84.2	66.1
SdAE (Chen et al., 2022c)	w/o	800	84.0	N/A
DDAE(ours)	w/o	800	**84.3**	67.7

In terms of linear probing, our approach surpasses the above MIM methods with the same training epochs, but is not as good as the contrastive-based methods. Contrastive-based methods compare across images while MIM-based methods exploit the whole image structure, which may care about more than 1000 classes (Chen et al., 2022b). This phenomenon is also reported in MAE (He et al., 2022) and BEiT (Bao et al., 2021), so for MIM-based methods, fine-tuning measurement may be a better metric to validate their effectiveness.

4.3 Analysis of DDAE

Migrating Deep Dynamic Supervision to Existing MIM methods. Although for simplicity we build our framework as a one-stage approach, our core design, Deep Dynamic Supervision(DDS), is compatible with other existing MIM methods. To demonstrate that, we conduct experiments to migrate DDS into representative one-stage and two-stage MIM methods, MAE (He et al., 2022) and BEiT-v2 (Peng et al., 2022a) respectively. We extract intermediate features of the encoder, then feed them into decoder together with the encoder's final output. The Reconstruction targets are all set the same as the final output. Note reconstruction losses for intermediate layers are multiplied by their corresponding dynamic loss weight as mentioned in Section 3.2. The plug-and-play results are reported in Table 2, which demonstrates our design's generality for both one-stage and two-stage approaches.

More local priors for ViTs. We conjecture that the fine-grained task built by MIM enables ViT to pay more attention to local information besides using global semantic inference, which inherently makes up for the absence of local prior in the structure of ViTs.

We compute averaged attention distance in each attention head of each layer (Dosovitskiy et al., 2020). We conduct experiments based on MAE since it is a simple enough MIM framework. The encoder is ViT-B with 100 epoch pretraining. Specifically, we draw vanilla MAE and DDS-MAE to demonstrate our design's effectiveness. As illustrated in Fig 4, DDS remarkably reduces the average attention distance of the model, especially in deep layers. This phenomenon indicates that through

Table 2: Plug-and-play Experiments.

Methods	Extra tokenizer	Pretraining Epochs	Finetune
MAE (He et al., 2022)	w/o	300	82.9
DDS-MAE	w/o	100	**83.2**
BEiT-v2 (Peng et al., 2022a)	w/	100	83.9
DDS-BEiT-v2	w/	100	**84.1**

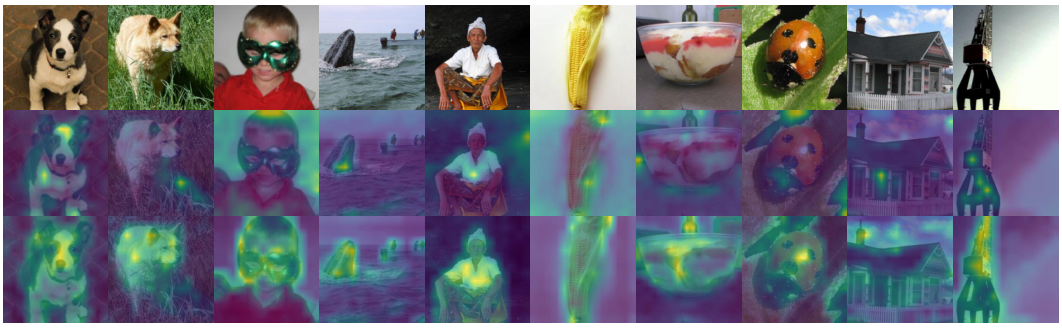
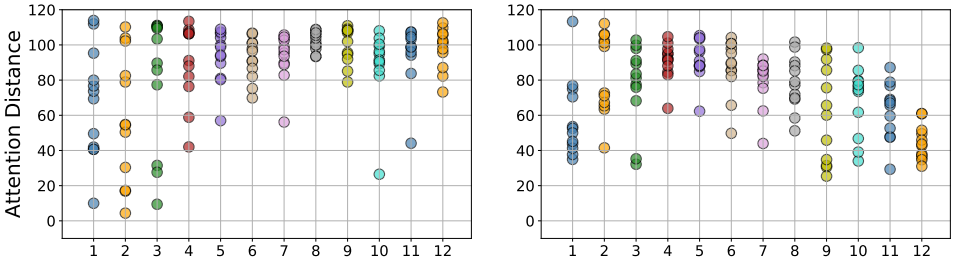


Figure 3: The attention map averaged over 12 attention heads between the class token and the patch tokens in the shallow layer (Block 3) of the ViT-B encoder pretrained on ImageNet-1K. Top: Input image, Middle: MAE, and Bottom: our DDAE. We can see that DDAE has formed effective target perception attention in the very early stage, while MAE still behaves to be rudimentary.



(a) MAE (b) DDS-MAE

Figure 4: The averaged attention distance in different attention heads (dots) w.r.t the layer number on MAE (a) and DDS-MAE (b) with ViT-B as the backbone.

deep dynamic supervision the model focuses more on local vicinity, thus obtaining more local priors, especially in deep layers. Moreover, for DDS part, different heads in relatively deeper layers behave more diversely, which may also contribute to DDS's effectiveness (+ 1.22% gains over MAE on ViT-S and more than + 0.3% gains on ViT-B).

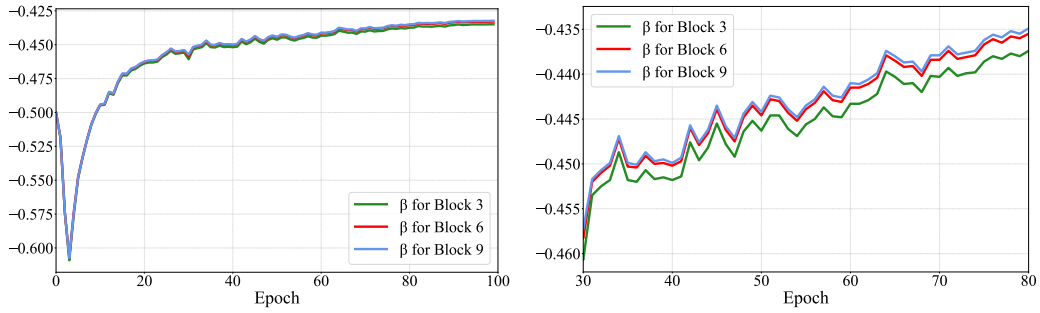
We further draw the curve of dynamic coefficients β changes during training for different layers in Fig 5. Firstly, for all intermediate layers, β drops sharply at the beginning, then rises slowly as the training progresses. Note that a smaller β value means that the model pays more attention to simple patches which are close to visible patches. Therefore, in the overall trend, each layer attaches more importance to the reconstruction of simple patches and gradually pays more attention to difficult ones. Secondly, for the comparison between layers, the shallow layer focuses more on simple patches, while the deep layer pays more attention to the difficult ones, which conforms to the characteristics of the model.

Ablation on Deep Dynamic Supervision. We conduct ablation studies to analyze the contributions of each component in Deep Dynamic Supervision. The models are finetuned on ImageNet-1K for 100 epochs. We set ViT-S as encoder and pretrain for 100 epochs. Table 3 shows that either ablation degrades the performance, suggesting that both deep supervision and dynamic loss benefit the training dynamics.

Ablation on Deep Self-Distillation. Then, we study Deep Self-Distillation in DDAE, we conduct ablation experiments with ViT-B as encoder, pretrain for 100 epochs and finetune 100 epochs on

Table 3: Ablation study on Deep Dynamic Supervision. Here D-MAE means applying Dynamic Loss to MAE, while DS-MAE means applying Deep Supervision to MAE. ViT-S is used as backbone.

Methods	**Deep Supervision**	**Dynamic Loss**	**Finetune**
MAE (He et al., 2022)			78.7
D-MAE		✓	78.9
DS-MAE	✓		79.7
DDS-MAE	✓	✓	**79.9**



(a) Overall trend (b) Detailed trend

Figure 5: β dynamic curve for intermediate layers. (a) is the overall trend for epoch 1 to epoch 100; (b) is the detailed trend for epoch 30 to epoch 80. Larger β means more importance is attached to difficult patches.

Table 4: Ablation study on Deep Self-Distillation. Here ✓means applying DDS on that part.

Methods	Pixel Regression	Feature Self-Distillation	Finetune
			83.0
		✓	83.2
	✓		83.3
DDAE	✓	✓	**83.5**

ImageNet-1K. Results in Table 4 suggests that applying DDS on Pixel and Feature both can improve the feature representation learned.

4.4 Semantic Segmentation

We evaluate the learned representation of DDAE on the ADE20K benchmark (Zhou et al., 2019), which consists of 25K images and 150 semantic categories. We use the UperNet (Xiao et al., 2018) task layer for semantic segmentation. We train Upernet 160K iterations with single-scale inference. The results are reported in Table 5 with mean Intersection of Union (mIoU) as the evaluation metric. Compared with previous state-of-the-art self-supervised models, our proposed DDAE outperforms all the one-stage methods, further validating the effectiveness of our framework.

Table 5: Semantic segmentation mIoU (%) comparison on ADE20K. ViT-B is used as backbone.

Methods	Extra Tokenizer	Pretraining Epochs	mIoU
BEiT (Bao et al., 2021)	w/	800	45.6
SimMIM (Xie et al., 2022b)	w/o	800	46.9
MAE (He et al., 2022)	w/o	1600	48.1
PeCo (Dong et al., 2021)	w/	800	48.5
CAE (Chen et al., 2022b)	w/	800	48.8
SdAE (Chen et al., 2022c)	w/o	800	49.0
BootMAE (Dong et al., 2022)	w/o	800	49.1
DDAE(ours)	w/o	800	**49.3**

5 Conclusion

In this paper, we propose a new framework DDAE with two core designs. (1) We present deep dynamic supervision to enable dynamic focus on different patches through pretraining, providing progressively richer representation. (2) We propose to self-distill intermediate features of momentum encoder with dynamic supervision, which enhance the representation from a high semantic level. More importantly, we find our approach can bring more local priors for ViTs, especially in deep layers, which we empirically find crucial for MIM's success.

REFERENCES

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. *arXiv preprint arXiv:2202.03555*, 2022.

Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

Josh Beal, Eric Kim, Eric Tzeng, Dong Huk Park, Andrew Zhai, and Dmitry Kislyuk. Toward transformer-based object detection. *arXiv preprint arXiv:2012.09958*, 2020.

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*, pp. 213–229. Springer, 2020.

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 IEEE/CVF International Conference on Computer Vision*, pp. 9650–9660, 2021.

Qiang Chen, Qiman Wu, Jian Wang, Qinghao Hu, Tao Hu, Errui Ding, Jian Cheng, and Jingdong Wang. Mixformer: Mixing features across windows and dimensions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5249–5259, 2022a.

Xiaokang Chen, Mingyu Ding, Xiaodi Wang, Ying Xin, Shentong Mo, Yunhao Wang, Shumin Han, Ping Luo, Gang Zeng, and Jingdong Wang. Context autoencoder for self-supervised representation learning. *arXiv preprint arXiv:2202.03026*, 2022b.

Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9640–9649, 2021.

Yabo Chen, Yuchen Liu, Dongsheng Jiang, Xiaopeng Zhang, Wenrui Dai, Hongkai Xiong, and Qi Tian. Sdae: Self-distillated masked autoencoder. *arXiv preprint arXiv:2208.00449*, 2022c.

Xiangxiang Chu, Zhi Tian, Yuqing Wang, Bo Zhang, Haibing Ren, Xiaolin Wei, Huaxia Xia, and Chunhua Shen. Twins: Revisiting the design of spatial attention in vision transformers. *Advances in Neural Information Processing Systems*, 34:9355–9366, 2021.

MMSegmentation Contributors. Mmsegmentation, an open source semantic segmentation toolbox, 2020.

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.

Xiaoyi Dong, Jianmin Bao, Ting Zhang, Dongdong Chen, Weiming Zhang, Lu Yuan, Dong Chen, Fang Wen, and Nenghai Yu. Peco: Perceptual codebook for bert pre-training of vision transformers. *arXiv preprint arXiv:2111.12710*, 2021.

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.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Alaaeldin El-Nouby, Gautier Izacard, Hugo Touvron, Ivan Laptev, Hervé Jegou, and Edouard Grave. Are large-scale datasets necessary for self-supervised pre-training? *arXiv preprint arXiv:2112.10740*, 2021.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16000–16009, 2022.

Gao Huang, Danlu Chen, Tianhong Li, Felix Wu, Laurens Van Der Maaten, and Kilian Q Weinberger. Multi-scale dense networks for resource efficient image classification. *arXiv preprint arXiv:1703.09844*, 2017.

Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 1725–1732, 2014.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Chen-Yu Lee, Saining Xie, Patrick Gallagher, Zhengyou Zhang, and Zhuowen Tu. Deeply-supervised nets. In *Artificial intelligence and statistics*, pp. 562–570. PMLR, 2015.

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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10012–10022, 2021.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Agata Mosinska, Pablo Marquez-Neila, Mateusz Koziński, and Pascal Fua. Beyond the pixel-wise loss for topology-aware delineation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3136–3145, 2018.

Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European conference on computer vision*, pp. 483–499. Springer, 2016.

Zhiliang Peng, Wei Huang, Shanzhi Gu, Lingxi Xie, Yaowei Wang, Jianbin Jiao, and Qixiang Ye. Conformer: Local features coupling global representations for visual recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 367–376, 2021.

Zhiliang Peng, Li Dong, Hangbo Bao, Qixiang Ye, and Furu Wei. Beit v2: Masked image modeling with vector-quantized visual tokenizers. *arXiv preprint arXiv:2208.06366*, 2022a.

Zhiliang Peng, Li Dong, Hangbo Bao, Qixiang Ye, and Furu Wei. Beit v2: Masked image modeling with vector-quantized visual tokenizers. *arXiv preprint arXiv:2208.06366*, 2022b.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pp. 8821–8831. PMLR, 2021.

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*, pp. 1–9, 2015.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pp. 10347–10357. PMLR, 2021a.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pp. 10347–10357. PMLR, 2021b.

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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 568–578, 2021.

Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14668–14678, 2022a.

Longhui Wei, Lingxi Xie, Wengang Zhou, Houqiang Li, and Qi Tian. Mvp: Multimodality-guided visual pre-training. *arXiv preprint arXiv:2203.05175*, 2022b.

Bichen Wu, Chenfeng Xu, Xiaoliang Dai, Alvin Wan, Peizhao Zhang, Zhicheng Yan, Masayoshi Tomizuka, Joseph Gonzalez, Kurt Keutzer, and Peter Vajda. Visual transformers: Token-based image representation and processing for computer vision. *arXiv preprint arXiv:2006.03677*, 2020.

Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. Cvt: Introducing convolutions to vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22–31, 2021.

Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 418–434, 2018.

Saining Xie and Zhuowen Tu. Holistically-nested edge detection. In *Proceedings of the IEEE international conference on computer vision*, pp. 1395–1403, 2015.

Zhenda Xie, Zigang Geng, Jingcheng Hu, Zheng Zhang, Han Hu, and Yue Cao. Revealing the dark secrets of masked image modeling. *arXiv preprint arXiv:2205.13543*, 2022a.

Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9653–9663, 2022b.

Baosong Yang, Longyue Wang, Derek Wong, Lidia S Chao, and Zhaopeng Tu. Convolutional self-attention networks. *arXiv preprint arXiv:1904.03107*, 2019.

Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zi-Hang Jiang, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 558–567, 2021.

Hanyuan Zhang, Hao Wu, Weiwei Sun, and Baihua Zheng. Deeptravel: a neural network based travel time estimation model with auxiliary supervision. *arXiv preprint arXiv:1802.02147*, 2018a.

Linfeng Zhang, Xin Chen, Junbo Zhang, Runpei Dong, and Kaisheng Ma. Contrastive deep supervision. *arXiv preprint arXiv:2207.05306*, 2022.

Muhan Zhang, Shali Jiang, Zhicheng Cui, Roman Garnett, and Yixin Chen. D-vae: A variational autoencoder for directed acyclic graphs. *Advances in Neural Information Processing Systems*, 32, 2019.

Zhenli Zhang, Xiangyu Zhang, Chao Peng, Xiangyang Xue, and Jian Sun. Exfuse: Enhancing feature fusion for semantic segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 269–284, 2018b.

Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6881–6890, 2021.

Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127(3):302–321, 2019.

Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020.

A TRAINING DETAILS

ImageNet Pretraining. We use Adam (Kingma & Ba, 2014) as the optimizer and train the DDAE for 100/800 epochs to illustrate both the rapid convergence and a higher upper limit. The learning rate is set as $1.5e^{-4} * batchsize/256$ with cosine learning rate decay and warmed up for 20/40 epochs respectively. The total batch size is set as 4096. The weight decay is set as 0.05. We increase the momentum parameter from 0.999 to 0.99999 in the first 400 epochs. The mask ratio is 75%.

ImageNet Fine-tuning. We follow the fine-tuning protocol in BEiT to use layer-wise learning rate decay as 0.75/0.65 for ViT-S/ViT-B respectively. The batch size is 1024, the warmup epoch is 5 and the weight decay is 0.05. For ViT-S, we train 100 epochs with a learning rate 1.6e-2. For ViT-B, We train 100 epochs with a learning rate 8e-3.

Linear probing. We use the LARS (Karpathy et al., 2014) optimizer with momentum 0.9. We train the model for 90 epochs with a batch size of 16384. The warmup epoch is 10 and the base learning rate is 0.1. We set weight decay as zero.

ADE20K Semantic segmentation. We use UperNet (Xiao et al., 2018) based on mmsegmentation (Contributors, 2020). Following BEiT, we use AdamW (Loshchilov & Hutter, 2017) optimizer with initial learning rate $4e^{-4}$, weight decay of 0.05 and batch size of 16 for $160K$ iterations. We warm up the learning rate for 1500 iterations and use a linear decay strategy. We set the layer decay as 0.65. Since ViTs can not output hierarchical features, we utilize four different scales FPNs to scale the feature map into different sizes. The input resolution is 512 × 512. We report single-scale test results.

B MORE ABLATION STUDIES

Ablation on the supervisory number K. We further explored the impact of supervisory number K on performance. Specifically, we set K to be 1, 2, 4, and 6 respectively. The method to determine the block index is the same as that described in Section 3.2. As shown in Table 6, the additional benefits brought by more than four supervisory signals have been saturated. So we set $K = 4$ as default in our framework.

Table 6: Ablation study on supervisory numbers K. Note all supervisions use Dynamic Loss.

K	**Block Index**	**Finetune**
1	$[12]$	78.9
2	$[6, 12]$	79.5
4	$[3, 6, 9, 12]$	**79.9**
6	$[2, 4, 6, 8, 10, 12]$	**79.9**

Ablation on the choice of β. We also explored the choice of β. Firstly, we fix β with different initialization, then we set β as learnable parameters. Table 7 demonstrates that learnable β benefits most.

Table 7: Ablation study on the choice of β. Fixed means β will not be updated in the pretraining.

β	**Initialization**	**Finetune**
Fixed	$[0, 0, 0, 0]$	79.7
Fixed	$[-0.5, -0.5, -0.5, -0.5]$	79.5
Learnable	$[0, 0, 0, 0]$	79.6
Learnable	$[-0.5, -0.6, -0.7, -0.8]$	79.8
Learnable	$[-0.5, -0.5, -0.5, -0.5]$	**79.9**

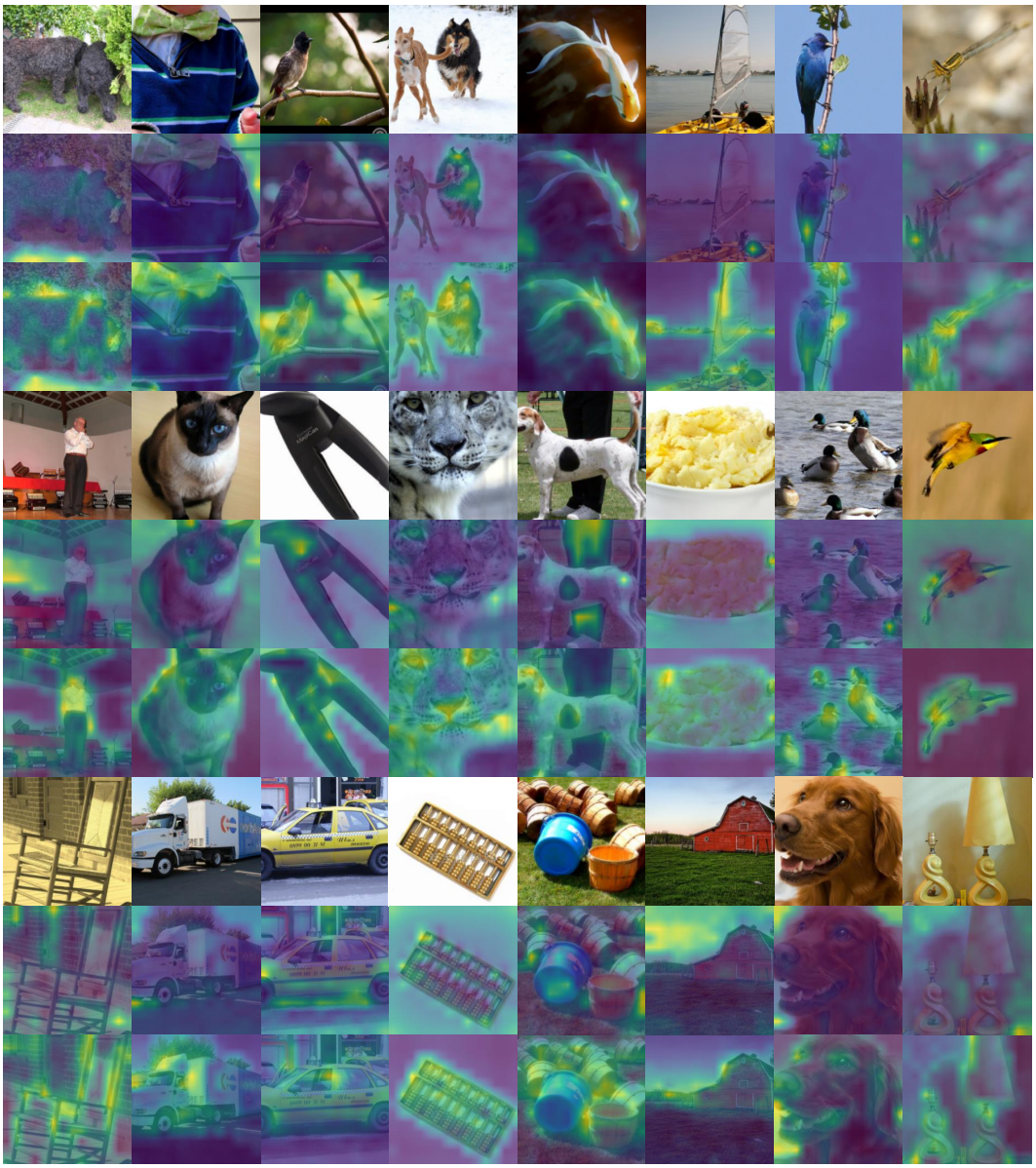


Figure 6: More attention map in the shallow layer (Block 3) of the ViT-B encoder of DDAE and MAE. The attention map is averaged over 12 attention heads between the class token and the patch tokens of the ViT-B encoder pretrained on ImageNet-1K. Top: Input image, Middle: MAE, and Bottom: our DDAE.

C More Visualizations

Our DDAE can form target-aware attention at a very early stage (Block 3), which is in sharp contrast to MAE. Here we show more shallow attention maps of DDAE and MAE in Fig. 6.