CoMA: Complementary Masking and Hierarchical Dynamic Multi-Window Self-Attention in a Unified Pre-training Framework

Jiaxuan Li[1 2], Qing Xu[1 2], Xiangjian He[1*], Ziyu Liu[1 2], Chang Xing[1], Zhen Chen[3], Daokun Zhang[1], Rong Qu[2], Chang Wen Chen[4]

[1] University of Nottingham Ningbo China
[2] University of Nottingham UK
[3] Yale University
[4] The Hong Kong Polytechnic University
{jiaxuan.li, qing.xu, sean.he, ziyu.liu, chang.xing, daokun.zhang}@nottingham.edu.cn,
{scxjl6, scxqx1, Ziyu.Liu1, rong.qu}@nottingham.ac.uk,
{zhen.chen}@yale.edu, {changwen.chen}@polyu.edu.hk

Abstract

Masked Autoencoders (MAE) achieve self-supervised learning of image representations by randomly removing a portion of visual tokens and reconstructing the original image as a pretext task, thereby significantly enhancing pretraining efficiency and yielding excellent adaptability across downstream tasks. However, MAE and other MAE-style paradigms that adopt random masking generally require more pre-training epochs to maintain adaptability. Meanwhile, ViT in MAE suffers from inefficient parameter use due to fixed spatial resolution across layers. To overcome these limitations, we propose the Complementary Masked Autoencoders (CoMA), which employ a complementary masking strategy to ensure uniform sampling across all pixels, thereby improving effective learning of all features and enhancing the model's adaptability. Furthermore, we introduce DyViT, a hierarchical vision transformer that employs a Dynamic Multi-Window Self-Attention (DM-MSA), significantly reducing the parameters and FLOPs while improving fine-grained feature learning. Pre-trained on ImageNet-1K with CoMA, DyViT matches the downstream performance of MAE using only 12% of the pre-training epochs, demonstrating more effective learning. It also attains a 10% reduction in pre-training time per epoch, further underscoring its superior pre-training efficiency.

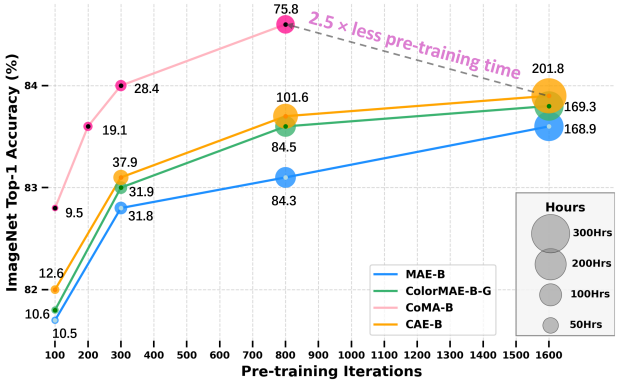

Figure 1: Visualization of the relationship between the number of pre-training iterations and ImageNet-1K classification accuracy for the base versions of MAE (He et al. 2022), CAE (Chen et al. 2024), ColorMAE (Hinojosa et al. 2024), and CoMA. The radius of each data point represents the total pre-training time, with larger dots indicating longer durations. Specific pre-training hours are annotated beside each point. All models were trained using a 60% masking ratio.

Introduction

In recent years, self-supervised learning has emerged as a prominent paradigm in computer vision research. It learns generalizable feature representations from unlabeled data by designing pretext tasks and demonstrates strong transferability over a wide range of downstream tasks (Ermolov et al. 2021; Oord, Li, and Vinyals 2018; Gui et al. 2024). Various self-supervised learning methods, including contrastive learning (Chen et al. 2020) and image reconstruction (Xie et al. 2022), have demonstrated performance comparable to or exceeding that of traditional supervised pre-training across a variety of downstream tasks such as image classification, object detection, and semantic segmentation, thus highlighting their significant potential for advancing visual representation learning.

Masked Autoencoders (MAE) have been extensively explored in recent studies on self-supervised learning, owing to their ability to model only visible tokens through a random masking strategy. This design substantially improves pretraining efficiency, producing a speed-up of 4× to 10× in the pre-training phase, while maintaining strong generalization performance across downstream tasks (He et al. 2022). Recent studies tend to guide the masking strategy through additional modules and implement random masking within the guided regions. (Shi et al. 2022; Chen et al. 2023; Krishna and Subramanyam 2025).

However, the use of random masking may result in an uneven allocation of sparse supervision signals, wherein certain regions are excessively sampled while others suffer from insufficient supervision. Moreover, introducing additional modules to guide the masking process inevitably in-

*Corresponding author.

Figure 2: Visualization of masking frequency over 1,600 iterations for Random masking (MAE) and Complementary Masking (CoMA). Dark red regions indicate patches with higher masking frequencies, while lighter regions correspond to less frequently masked patches. Both strategies employ a masking ratio of 60%.

curs increased computational overhead during pre-training. In practice, such sparse supervision signals necessitate additional resampling iterations to ensure that the sampling frequencies of features converge toward statistical uniformity. As shown in Fig. 2, the sampling statistics of MAE's random masking over 1,600 iterations reveal that not all patches are uniformly selected as masked tokens, even after extensive pre-training. This uneven distribution may impair the model's ability to learn from regions with high masking frequencies, and under such conditions, the model may prematurely converge falsely assuming that optimal learning has been achieved under the current masking strategy. Moreover, MAE adopts ViT as its backbone, which partitions images into independent patches and models long-range dependencies. However, this design overlooks local structures and internal correlations within patches, and its fixed spatial resolution further limits adaptability, resulting in low parameter efficiency (Yuan et al. 2021; Ryali et al. 2023; Liu et al. 2021; Li et al. 2025).

To enhance both the efficiency and the representational quality of pre-training, we introduce Complementary Masked Autoencoder (CoMA). This approach generates a pair of complementary random masks to ensure that all visual tokens have an equal probability of being selected as visible or masked in each iteration. Based on this, the two complementary sets of visible tokens are fed into the adaptive model and the evaluation model, respectively, which share an identical architecture. Importantly, the parameters of the evaluation model are entirely frozen and updated solely through the guidance of the adaptive model, introducing negligible overhead to pre-training efficiency. This design facilitates uniform feature sampling as shown in Fig. 2, mitigates the risk of convergence to suboptimal solutions caused by imbalanced masking strategies, ultimately enhancing the model's adaptability in downstream tasks. Meanwhile, we design a lightweight hierarchical Vision Transformer, DyViT, which integrates the proposed Dynamic Multi-Window Self-Attention to model the relationship between feature points and multi-scale representations, thereby enhancing the perceptual ability of each layer. DyViT significantly reduces the number of parameters and FLOPs, thereby improving the pre-training efficiency of masked autoencoders. We have pre-trained DyViT on ImageNet-1K using the Complementary Masked Autoencoder (CoMA) approach. It converges in fewer epochs and achieves approximately 10% faster pre-training speed than MAE as shown in Fig. 1. CoMA enables more efficient parameter utilization, facilitates faster convergence, and reduces the model's dependence on extended pre-training durations. We have evaluated DyViT on multiple downstream tasks, where it demonstrates highly competitive performance. Our main contributions are summarized as follows:

- We construct a complementary masking strategy to ensure that all patches receive effective supervision in each pre-training iteration. This approach improves data utilization and mitigates the risk that the model is converging to a local optimum.
- We propose DyViT, which adaptively captures multi-scale information to enhance downstream performance. Its hierarchical architecture effectively improves the parameter efficiency of the Vision Transformer.
- The proposed Dynamic Multi-Window Self-Attention enables multi-scale modeling within ViT, fundamentally enhancing the model's ability to perceive features at varying granularities.
- CoMA pre-training enables DyViT to achieve superior downstream performance with shorten pre-training time, highlighting its efficiency and effectiveness.

Related Work

Masked Autoencoders Representation Learning

Masked Autoencoders (MAE) draw inspiration from masked language modeling (MLM) in natural language processing, where a portion of input tokens is masked and predicted using the surrounding context (Devlin et al. 2019). Unlike NLP, where the number of tokens does not affect the computation of self-attention scores due to the inherent design of vision transformers, MAE removes masked patches directly from the input sequence during encoding (He et al. 2022; Feichtenhofer et al. 2022). Some recent studies have proposed replacing the random masking strategy in MAE to enhance downstream performance. These approaches mainly involve introducing additional modules (Shin et al. 2024; Madan et al. 2024; Wang et al. 2024), or using a teacher-student framework (Zhu et al. 2024; Mo 2025), which encourages Vision Transformer to focus on regions with high semantic or visual information density during the reconstruction task. Although constraining the masking region using image-informed strategies can improve focus, it still suffers from uneven sampling across patches, often requiring longer training to converge. Additionally, such methods increase models' complexity and computational costs. Enhancing pre-training efficiency while maintaining robustness remains a key challenge.

Multi-scale Perception of Vision Transformers

Multi-scale Vision Transformers effectively capture low-level visual information in the shallow layers and progressively extract more complex semantic features in the deeper

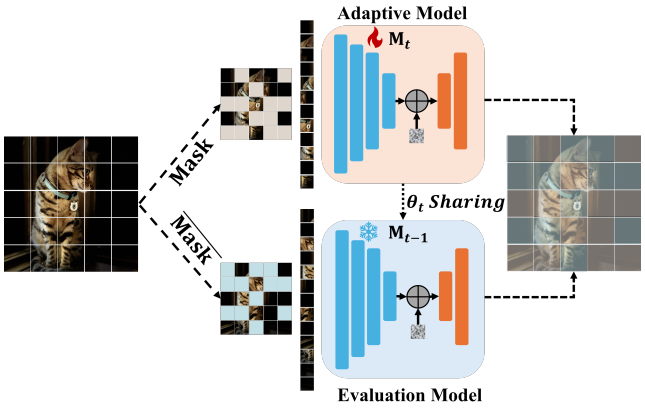

Figure 3: **CoMA: Complementary Masked Autoencoder.** Model M_t serves as the adaptive model and participates in gradient backpropagation, while model M_{t-1} acts as the evaluation model and remains completely frozen. The parameters of M_{t-1} are updated solely based on those of M_t at time step t.

layers (Fan et al. 2021). The hierarchical design effectively enhances the model's ability to capture objects of varying sizes within an image, while encouraging the model to maintain robust recognition capability under scale-equivariant conditions (Tian et al. 2023). Some recent arts demonstrate that the ingenious hierarchical design remains highly effective in Vision Transformers (Liu et al. 2024; Ghahremani et al. 2024). For instance, the Pyramid Vision Transformer (PVT) (Wang et al. 2021) employs a hierarchical pyramid architecture to generate multiscale feature representations, thus enhancing the model's representational capacity. Empirical results validate that this hierarchical design facilitates accurate dense predictions for high-resolution inputs in various vision tasks. The Swin Transformer (Liu et al. 2021, 2022a) introduces a sliding window mechanism that confines self-attention computation to local windows, thus reducing computational complexity. Hiera, proposed in (Ryali et al. 2023), improves downstream task performance through simple yet effective modeling of unit attention, enhancing the model's ability to perceive local features. Prior studies (Long, Shelhamer, and Darrell 2015; Liu et al. 2016; Lin et al. 2017; Zhao et al. 2017; Wu et al. 2021; Zhang et al. 2021; Ghahremani et al. 2024) consistently demonstrate that aggregating information across different granularities significantly boosts model performance on downstream tasks.

Methods

The MAE pre-training process relies on random masking, which results in uneven mask coverage that limits the model's learning capacity. This misleads the model into believing it has reached an optimal solution, leading to convergence on suboptimal representations derived from insufficient random sampling. Furthermore, the fixed-resolution design of the ViT used in MAE hinders its ability to capture fine-grained features. To address these issues, we introduce CoMA, a novel pre-training framework, alongside DyViT, a hierarchical Vision Transformer, to enhance pre-training efficiency and effectiveness.

Complementary Masked Autoencoders

We propose the Complementary Masked Autoencoder, which employs complementary masking strategies to ensure that the model can predict two mutually exclusive masks with only a single parameter update. This design also maintains the pre-training efficiency.

Complementary Masking. CoMA adopts a dual-branch complementary masking strategy to ensure that all patches have equal chances of being masked and unmasked within a single epoch, as illustrated in Fig. 3, thus promoting more comprehensive feature learning.

In the two branches, we adopt a complementary random masking strategy to ensure that different subsets of patches are visible in each branch, such that the two masks satisfy the condition $\text{Mask} + \overline{\text{Mask}} = \mathbf{1}$, where Mask denotes the masking strategy used by the adaptive model and $\overline{\text{Mask}}$ corresponds to that of the evaluation model. Note that both masks are binary matrices, with 1 indicating the positions of preserved vision tokens and 0 representing the tokens to be removed. Given any input image $\mathbf{X}$, the feature processing proceeds as follows:

$$\mathrm{X}^{\mathrm{M}} = \mathrm{X} \circ \text{Mask}, \quad \mathrm{X}^{\mathrm{C}} = \mathrm{X} \circ \overline{\text{Mask}}, \tag{1}$$

where $\circ$ denotes the Hadamard product. X^{M} denotes the visible tokens retained by Mask, while X^{C} represents those retained by its complement $\overline{\text{Mask}}$. If the spatial dimensions of Mask, $\overline{\text{Mask}}$, and X are inconsistent, interpolation is applied to masks ensure spatial alignment. We remove all tokens with a value of zero and feed the remaining tokens into the encoder. At the end of the encoder, we reinsert empty patches at the positions masked out (i.e., those with zeros in Mask and $\overline{\text{Mask}}$) to restore the original input structure, following the same procedure as MAE. Finally, the reconstructed features are passed to the decoder, producing $\mathrm{X}^{\mathrm{M}}_{\mathrm{rec}}$ and $\mathrm{X}^{\mathrm{C}}_{\mathrm{rec}}$, respectively. Subsequently, we assemble the reconstructed patches into a complete image, which corresponds to a full-image reconstruction. The operation is performed by:

$$\mathrm{X}_{\mathrm{rec}} = \mathrm{X}^{\mathrm{M}}_{\mathrm{rec}} \circ \overline{\text{Mask}} + \mathrm{X}^{\mathrm{C}}_{\mathrm{rec}} \circ \text{Mask}, \tag{2}$$

where $\mathrm{X}_{\mathrm{rec}}$ is the final reconstructed features with the same shape as X. The final reconstructed full images are supervised using the Mean Squared Error (MSE) loss, formulated as:

$$\mathcal{L}_{\mathrm{MSE}} = \|\mathrm{X}_{\mathrm{rec}} - \mathrm{X}\|_F^2. \tag{3}$$

Efficient Pre-training. In CoMA, the parameters of the evaluation model $\theta_{\text{Evaluation}}$ are frozen to improve the pre-training efficiency. These parameters are directly updated from the adpative model θ_{Adaptive} in the previous pre-training step. This strategy, inspired by the Double Deep Q-Network (Van Hasselt, Guez, and Silver 2016), enhances the stability of pre-training. As training progresses, the predictions of

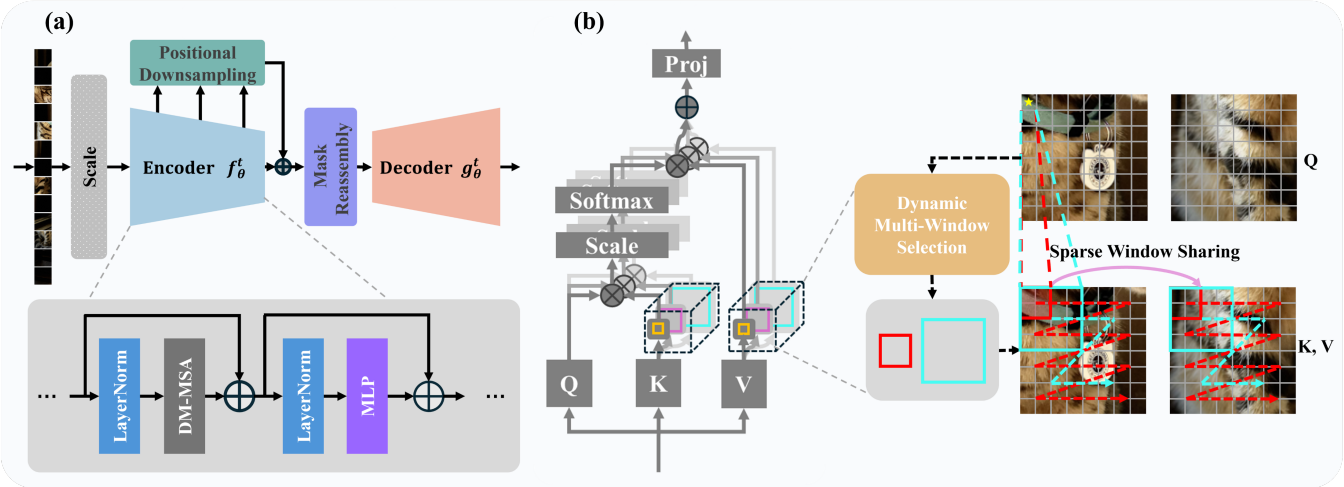

Figure 4: The structure of the proposed DyViT model and its core attention mechanism. (a) Overview of the DyViT architecture. (b) Illustration of the Dynamic Multi-window Self-Attention module (DM-MSA).

$\theta_{\text{Evaluation}}$ are expected to approximate those of θ_{Adaptive}. The parameter update is performed as:

$$\theta^{(t)}_{\text{Evaluation}} \leftarrow \theta^{(t-1)}_{\text{Adaptive}}. \tag{4}$$

We employ parameter sharing "$\leftarrow$" instead of an EMA strategy (Tarvainen and Valpola 2017) because the Evaluation Model functions merely as an auxiliary branch that acts as a shadow of the Adaptive model to assess its performance under a complementary masking scheme. This pre-training strategy enables dense supervision over all patches within the image, effectively alleviating the sampling bias introduced by random masking and ensuring more uniform coverage across the spatial domain.

We freeze the Evaluation Model for two primary reasons. Firstly, we aim to improve the efficiency of pre-training. Compared with contrastive learning (Chen et al. 2020), only the Adaptive model needs to be trained after freezing the parameters, which significantly reduces the computational cost. Secondly, because the Evaluation Model acts solely as a projection of the Adaptive Model and is responsible only for predicting the Adaptive Model's output under the complementary mask, the temporal offset between them contributes to the stability of the improved model. This mechanism aligns with the principles of Double Deep Q-Network (Van Hasselt, Guez, and Silver 2016). Moreover, it enables parallel inference on the input data, avoiding the inefficiency of sequential reprediction.

Hierarchical Dynamic Vision Transformer

Hierarchical Vision Transformers progressively increase channel depth while reducing spatial resolution to capture high-level semantics. This structure improves parameter efficiency, enables multi-scale feature extraction, and reduces overall complexity. Building upon this idea, we design DyViT, a framework that can be seamlessly integrated into the CoMA pre-training paradigm and other MAE variants.

DyViT consists of four layers, with spatial resolutions progressively downsampled by factors of 4×, 8×, 16×, and 32× relative to the original input size. In the *scale* layer, we apply a convolution operation with a kernel size of 7×7, stride of 4×4 and padding of 3×3 to downsample the feature maps by a factor of $4\times$. We then apply patch-level masking on the feature maps, consistent with Equation (1), followed by the removal of the masked patches. All other layers employ a 2×2 max pooling operation within each patch to achieve the reduction in spatial resolution. In the first two layers, we employ multiple sets of DM-MSA blocks to model spatial correlations between pixels and multi-scale feature descriptors extracted with different window sizes, as shown in Fig. 4(a). This not only reduces the number of model parameters, but also facilitates effective modeling of spatial relationships across scales. In the deeper layers (i.e., layers 3 and 4), we adopt global self-attention (Dosovitskiy et al. 2020) to capture long-range dependencies and enhance high-level semantic understanding.

As illustrated in Fig. 4(a), the *Positional Downsampling* module is designed to extract features from the outputs of different layers and fuse them together to capture multi-scale information. The output features X_i from the four layers input to this module have the following shapes:

$$\mathrm{X}_i \in \mathbb{R}^{B \times fnp_i^2 \times C_i},\ i \in \{1,2,3,4\},\ p_i \in \{8,4,2,1\}, \tag{5}$$

where B is the batch size, n the number of patches, f the masking ratio, and C_i the channel dimension.

For the output features X_i from four layers, we perform patch downsampling using strided convolution. The detailed operation is as:

$$\mathrm{Y}_i = \mathrm{Conv}_{k,s}(\mathrm{Y}_{i-1}) + \mathrm{X}_{i+1},\quad i \in \{1,2,3\}, \tag{6}$$

where Y_{i-1} is initialized as X_1, and both the kernel size k and the stride s are set to 2^{4-i}. After hierarchical fusion, the output feature $\mathrm{X}_{\text{out}} = \mathrm{Y}_3$ is obtained, which has the

shape $\mathbb{R}^{B\times n\times C_4}$. The fused representation is subsequently processed by the *Mask Reassembly* module. Specifically, noise patches are inserted at the positions corresponding to the masked regions. The resulting latent features are then passed to a decoder composed of 8× global attention blocks (Dosovitskiy et al. 2020) for reconstruction.

Dynamic Multi-Window Self-Attention

Our design facilitates attention interactions between individual feature points and region-level features extracted from multiple window scales. Importantly, our attention design allows information exchange across patches, thereby enhancing global context modeling at shallow layers while maintaining computational efficiency during both training and inference.

Building upon the strength of self-attention in capturing global dependencies, we apply the standard query, key and value projections to the input feature X in the encoder. For any input $\mathrm{X} \in \mathbb{R}^{B\times np^2\times C}$, the initial projection of features into the query (Q), key (K), and value (V) representations is defined by:

$$\mathrm{Q} = \mathrm{X}\mathrm{W}^{Q}, \quad \mathrm{K} = \mathrm{X}\mathrm{W}^{K}, \quad \mathrm{V} = \mathrm{X}\mathrm{W}^{V}, \tag{7}$$

where $\mathrm{W}^{Q}, \mathrm{W}^{K}, \mathrm{W}^{V} \in \mathbb{R}^{C\times C}$. To capture multi-scale information at a specific resolution, our Dynamic Multi-Window Selection module first determines candidate convolutional kernel sizes, strides, and the number of branches based on the patch size as shown in Fig. 4(b). The operation of the Dynamic Multi-Window Selection module can be formulated as:

$$\mathcal{K} = \left\{ k_i \;\middle|\; k_i = \frac{p}{2^i},\ i \in \mathbb{N},\ 2^i \mid p \right\}, \tag{8}$$

where $k_i \in \mathcal{K}$ denotes the i-th possible kernel size and p is the patch size. Based on each kernel size in set $\mathcal{K}$, we apply stride convolution to extract features in a sparse manner. As illustrated in Fig. 4(b), the convolutional kernel operates strictly within the local region of each patch, capturing localized information. To enable the model to capture multi-scale features, we further incorporate a self-attention mechanism to model spatial relationships across positions and scales. The operation is formulated as:

$$\mathrm{Attn}_k(\mathrm{Q},\mathrm{K},\mathrm{V}) = \mathrm{Softmax}\left(\frac{\mathrm{Q}\left(\mathrm{Conv}_k(\mathrm{K})\right)^{\top}}{\sqrt{d}}\right)\mathrm{Conv}_k(\mathrm{V}), \tag{9}$$

where $\mathrm{Conv}_k(\cdot)$ denotes a strided convolution with both kernel size and stride equal to $k \in \mathcal{K}$. Due to the inherently sparse sampling nature of strided convolution, a single-scale window lacks the capacity to capture correlations between adjacent features. In contrast, a multi-scale strided convolutional design effectively addresses this limitation. Larger kernels offer broader receptive fields that subsume those of smaller ones, enabling the recovery of local feature correlations that may be overlooked at finer scales. Finally, the multi-scale features are integrated and linearly projected to enhance the model's ability to capture information over different levels of granularity. The procedure is as:

$$\mathrm{X}_{out} = \left[\sum_{k\in\mathcal{K}} \mathrm{Attn}_k(\mathrm{Q},\mathrm{K},\mathrm{V})\right]\mathrm{W}_{out}, \tag{10}$$

where W_{out} denotes the linear projection with shape $\mathbb{R}^{C\times C}$.

Experimental Results

Implementation Details

Model	#Channels	#Blocks	#Heads
DyViT-S	[96-192-384-768]	[1-2-11-2]	[2-4-8-16]
DyViT-B	[112-224-448-896]	[2-3-16-3]	[2-4-8-16]

Table 1: DyViT Configurations. #Channels represents the number of channels in each layer, #Blocks denotes the number of Transformer blocks per layer, and #Heads refers to the number of attention heads in the multi-head attention mechanism for each layer.

We perform pre-training on the ImageNet-1K training set using 4× H200 GPUs, each equipped with 141 GB of memory. The model is pre-trained with a batch size of 4096 using a patch size of 32×32 and a masking ratio of 60% for the Adaptive model and 40% for the Evaluation model. On average, every 100 epochs take approximately 9.5 hours for DyViT-B. AdamW is employed as the optimizer with a batch size of 4096. For downstream tasks with resolutions differing from pre-training, positional embeddings are reinserted prior to fine-tuning. Only the encoder weights are transferred. In image classification, global average pooling followed by a single-layer MLP maps the features to the target label space. The configurations of DyViT-Small and DyViT-Base are detailed in Table 1.

Model Evaluation

Image Classification. We present a comprehensive performance comparison among self-supervised Vision Transformers, as summarized in Table 2. Leveraging the CoMA pre-training framework, DyViT consistently outperforms a range of competitive architectures. Remarkably, when compared to the MAE's pre-trained Vision Transformer, DyViT pre-trained under CoMA achieves superior results using 300 epochs, highlighting its exceptional pre-training efficiency.

To further substantiate this observation, we conduct a direct comparison between DyViT models pre-trained with MAE and CoMA. At 800 training epochs, the DyViT model pre-trained with CoMA outperforms MAE's pre-trained counterpart by +0.4, reaffirming the superior pre-training effectiveness afforded by the CoMA framework. Notably, under the same pre-training budget, CoMA's pre-trained DyViT attains a top-1 classification accuracy of 84.6%, surpassing numerous existing state-of-the-art approaches. These findings collectively underscore the significant advantages of CoMA in enhancing effectiveness of the pre-training process.

Moreover, DyViT exhibits notable architectural efficiency. Compared to ViT-B operating at the same input

PT-Method	PT-Task	Arch.	Epochs	Acc.	FLOPs	Params
DINO (Caron et al. 2021)	CL	ViT-S	3200	82.0	5	22
BEiT (Bao et al. 2021)	MIM	ViT-S	800	81.4‡	5	22
MAE (He et al. 2022)	MAE	ViT-S	800	81.6‡	5	22
SimMiM (Xie et al. 2022)	MIM	Swin-S	1600	83.2	9	50
iBOT (Zhou et al. 2022)	MIM+CL	ViT-S	3200	82.3	5	22
DyViT	CoMA	DyViT-S	300	83.6	6	35
DyViT	CoMA	DyViT-S	800	**83.9**	6	35
DINO (Caron et al. 2021)	CL	ViT-B	1600	83.6	18	87
BEiT (Bao et al. 2021)	MIM	ViT-B	800	83.2	18	87
MAE (He et al. 2022)	MAE	ViT-B	1600	83.6	18	87
iBOT (Zhou et al. 2022)	MIM+CL	ViT-B	1600	84.0	18	87
SimMiM (Xie et al. 2022)	MIM	Swin-B	1600	83.8	15	88
A^2MIM (Li et al. 2022b)	MIM	ViT-B	800	84.3	18	87
Hybrid Distill (Shi et al. 2023)	Distill-MAE	ViT-B	300	83.7	18	87
CAE (Chen et al. 2024)	MIM	ViT-B	1600	83.9	18	87
ColorMAE (Hinojosa et al. 2024)	MAE	ViT-B	1600	83.8	18	87
DeepMIM-MAE (Ren et al. 2025)	Deep-supervision + MAE	ViT-B	1600	84.0	18	87
DyViT	MAE	DyViT-B	300	83.6	12	70
DyViT	CoMA	DyViT-B	300	83.9	12	70
DyViT	MAE	DyViT-B	800	84.2	12	70
DyViT	CoMA	DyViT-B	800	**84.6**	12	70

Table 2: Comparison of Top-1 Accuracy (%) between DyViT and state-of-the-art models on Imagenet-1K dataset. All models are evaluated with a resolution of 224×224 under consistent downstream configurations. PT-Task refers to the pre-training methods. Bold indicates the best performance. Results marked with ‡ are sourced from (Zhou et al. 2022).

resolution, DyViT-B achieves a reduction of approximately 20% in the number of parameters and a decrease of 33% in FLOPs. Despite the reductions in both number of parameterss and FLOPs, DyViT achieves superior performance, highlighting its effective parameter utilization.

Semantic Segmentation. For semantic segmentation on the ADE20K dataset (Zhou et al. 2017), DyViT follows the same fine-tuning protocol as MAE, employing UperNet as the decoder (Xiao et al. 2018). Table 3 summarizes the performance of DyViT compared to state-of-the-arts approaches on this benchmark. Remarkably, our model achieves 51.5 mIoU with only 800 pre-training epochs, significantly outperforming existing state-of-the-art (SOTA) methods, and outperforms MAE and ColorMAE by 3.4% and 2.2%, respectively, on the mIoU metric. These results strongly validate the effectiveness of the proposed hierarchical design in improving the model's ability to capture fine-grained semantic representations. Furthermore, DyViT captures visual tokens at different scales with a dynamic multi-window strategy, modeling and fusing features extracted from keypoints and windows at multiple scales to enhance perception across multiple granularities.

Object Detection and Instance Segmentation. We fine-tune Mask R-CNN (He et al. 2017) in an end-to-end manner on the COCO dataset (Lin et al. 2014) and report both AP^{box} for object detection and AP^{mask} for instance segmentation. Our implementation follows the ViT-Det training methods (Li et al. 2022c), while integrating multi-scale features extracted by DyViT into a Feature Pyramid Network (FPN) (Lin et al. 2017). As shown in Table 3, our model surpasses the SOTA method, Hiera-B (Ryali et al. 2023), by 0.9 in AP^{box} and 1.2 in AP^{mask}. Compared to MAE (He et al. 2022), our approach achieves a notable improvement of 2.8 in AP^{box} and 2.6 in AP^{mask}. DyViT's hierarchical structure and DM-MSA's coarse-to-fine modeling enable more robust object detection and instance segmentation.

PT-Task	Epochs	ADE20K	COCO	
		mIoU	AP^{box}	AP^{mask}
DINO (Caron et al. 2021)	400	47.2	46.8	41.5
SdAE (Chen et al. 2022)	300	48.6	48.9	43.0
MAE (He et al. 2022)	1600	48.1	50.3	44.9
A^2MIM (Li et al. 2022b)	800	49.0	49.4	43.5
Hiera (Ryali et al. 2023)	1600	50.8	52.2	46.3
ColorMAE (Hinojosa et al. 2024)	1600	49.3	50.1	44.4
CAE (Chen et al. 2024)	1600	50.2	50.2	44.2
CoMA	800	**51.5**	**53.1**	**46.5**

Table 3: Performance comparison on ADE20K (Zhou et al. 2017) for semantic segmentation and COCO (Lin et al. 2014) for object detection and instance segmentation. All models are pre-trained on ImageNet-1K and use the base version. Input resolutions are 512×512 for ADE20K (Zhou et al. 2017) and 768×768 for COCO with Mask R-CNN (He et al. 2017).

Ablation Study

Effectiveness of CoMA. We evaluate the image classification performance of masked autoencoder frameworks, as presented in Table 4. To ensure a fair comparison, ViT-B is adopted as the unified backbone (Dosovitskiy et al. 2020). According to the evaluation results, ViT-B pre-trained with

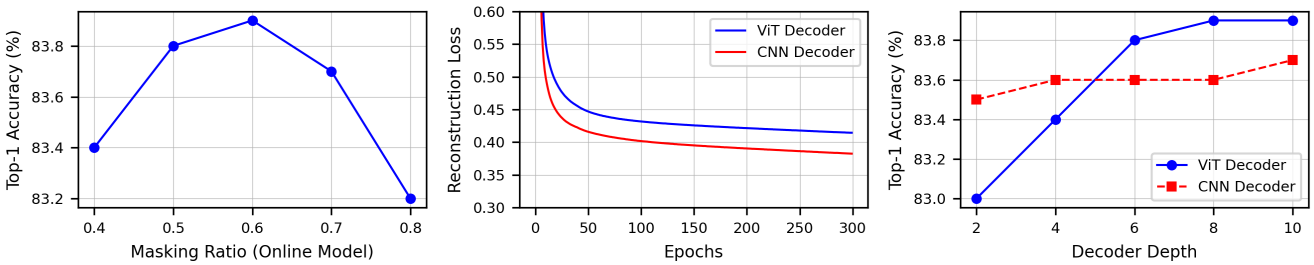

Figure 5: Ablation study across masking ratios with 32×32 patches and 300 pre-training epochs (left); reconstruction loss on the validation set using an 8-layer decoder (middle); impact of decoder depth on classification transferability (right).

Method	Epochs	Top-1 Acc.
BEiT (Bao et al. 2021)	800	83.2
MAE (He et al. 2022)	1600	83.6
MultiMAE (Bachmann et al. 2022)	1600	83.3
SemMAE (Li et al. 2022a)	800	83.3
ColorMAE (Hinojosa et al. 2024)	1600	83.8
CoMA (Ours)	800	**84.1**

Table 4: Comparison of CoMA with other masked modeling methods using ViT-Base. All methods are pre-trained and fine-tuned on ImageNet-1K with 224×224 resolution.

the CoMA framework achieves a top-1 accuracy of 84.1% on ImageNet-1K, surpassing other MAE-based methods with the same or fewer pre-training epochs. This shows that CoMA facilitates faster model convergence. CoMA's complementary masking strategy significantly enhances pre-training efficiency by enabling more informative and comprehensive sampling across all patches, thereby improving data utilization and effectively reducing the overall pre-training duration.

Masking Ratio. In MAE (He et al. 2022), the 16×16 patch division benefits high masking ratios, such as 75%, by allowing for finer learning of local textures and structures. Smaller patches promote more dispersed visible patches, enhancing the encoder's capacity with diverse contextual information. However, compared to 16×16, a patch size of 32×32 can improve pre-training efficiency more effectively, as it results in fewer tokens to process. Meanwhile, since DyViT applies a $4\times$ downsampling in the first layer to filter out redundant information, using larger patches is more suitable for our architecture, ensuring that each patch contains sufficient information to support the modeling of internal features through Dynamic Multi-Window Self-Attention.

Setting 300 epochs as the baseline, we systematically evaluate the impact of different masking ratios, with the results presented in Fig. 5. Empirical results indicate that a masking ratio of 60% yields the best performance for the adpative model. In contrast, excessively high masking ratios lead to a notable decline in accuracy. As larger patches encapsulate more information, the corresponding masked regions obscure significantly more features, covering up to four times the content of a patch of 16×16. Consequently, excessive masking impairs semantic understanding and learning efficiency. Our model's multi-scale feature modeling within each patch effectively mitigates limitations of larger patch sizes.

Decoder Depth. We evaluate two decoder architectures: ConvNeXt (Liu et al. 2022b) and ViT (Dosovitskiy et al. 2020). The results show that the ConvNeXt decoder achieves superior reconstruction performance on the ImageNet-1K validation set, demonstrating lower loss values and faster convergence (see Fig. 5). This advantage is primarily attributed to the stronger inductive bias inherent in CNN-based decoders.

However, reconstruction loss alone does not fully reflect the generalization capacity of the pre-trained model. In downstream ImageNet-1K classification tasks, increasing the decoder depth to 6 layers enables DyViT pre-trained with a ViT decoder to outperform its counterpart pre-trained with a ConvNeXt decoder. Moreover, the performance of the ViT decoder tends to converge between 8 and 10 layers. Based on these observations, we adopt a decoder composed of 8 ViT blocks in our model.

Pre-training Efficiency. We conduct pre-training on DyViT, MAE (He et al. 2022), CAE (Chen et al. 2024), and ColorMAE (Hinojosa et al. 2024), and compare their pre-training efficiency under the same masking ratio, as shown in Fig. 1. Our model demonstrates superior pre-training efficiency compared to state-of-the-art methods, achieving approximately 10% faster training speed than MAE (He et al. 2022) and ColorMAE (Hinojosa et al. 2024), while requiring only half the number of pre-training epochs to significantly surpass their performance. In particular, it outperforms CAE's 1600-epoch results with only 800 epochs. In downstream tasks, it achieves +2.9 AP^{box} over CAE (Chen et al. 2024) in object detection and segmentation (Table 3), and improves top-1 accuracy by +1.0 and +0.8 over MAE and ColorMAE, respectively, in ImageNet-1K classification, while using fewer parameters and FLOPs.

Conclusion

Efficient and effective masked autoencoders are increasingly essential for better utilization of limited computational resources, since principled and adaptive masking strategies can significantly reduce pre-training time and improve downstream task performance. In this work, we propose

CoMA, which provides richer complementary information to enhance representation learning, alongside DyViT, an efficient multi-scale vision transformer that improves parameter efficiency and accelerates pre-training, while maintaining strong performance on diverse complex natural image tasks. We hope that our exploration will inspire new perspectives and contribute to future advances in the vision community.

References

Bachmann, R.; Mizrahi, D.; Atanov, A.; and Zamir, A. 2022. Multimae: Multi-modal multi-task masked autoencoders. In *European Conference on Computer Vision*, 348–367. Springer.

Bao, H.; Dong, L.; Piao, S.; and Wei, F. 2021. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*.

Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; and Joulin, A. 2021. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, 9650–9660.

Chen, H.; Zhang, W.; Wang, Y.; and Yang, X. 2023. Improving masked autoencoders by learning where to mask. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, 377–390. Springer.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, 1597–1607. PmLR.

Chen, X.; Ding, M.; Wang, X.; Xin, Y.; Mo, S.; Wang, Y.; Han, S.; Luo, P.; Zeng, G.; and Wang, J. 2024. Context autoencoder for self-supervised representation learning. *International Journal of Computer Vision*, 132(1): 208–223.

Chen, Y.; Liu, Y.; Jiang, D.; Zhang, X.; Dai, W.; Xiong, H.; and Tian, Q. 2022. Sdae: Self-distillated masked autoencoder. In *European conference on computer vision*, 108–124. Springer.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 4171–4186.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Ermolov, A.; Siarohin, A.; Sangineto, E.; and Sebe, N. 2021. Whitening for self-supervised representation learning. In *International conference on machine learning*, 3015–3024. PMLR.

Fan, H.; Xiong, B.; Mangalam, K.; Li, Y.; Yan, Z.; Malik, J.; and Feichtenhofer, C. 2021. Multiscale vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, 6824–6835.

Feichtenhofer, C.; Li, Y.; He, K.; et al. 2022. Masked autoencoders as spatiotemporal learners. *Advances in neural information processing systems*, 35: 35946–35958.

Ghahremani, M.; Khateri, M.; Jian, B.; Wiestler, B.; Adeli, E.; and Wachinger, C. 2024. H-vit: A hierarchical vision transformer for deformable image registration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11513–11523.

Gui, J.; Chen, T.; Zhang, J.; Cao, Q.; Sun, Z.; Luo, H.; and Tao, D. 2024. A survey on self-supervised learning: Algorithms, applications, and future trends. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12): 9052–9071.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16000–16009.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, 2961–2969.

Hinojosa, C.; et al. 2024. ColorMAE: Exploring data-independent masking strategies in Masked AutoEncoders. In *European Conference on Computer Vision*, 432–449. Springer.

Krishna, M.; and Subramanyam, A. 2025. Keypoint aware masked image modelling. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.

Li, G.; Zheng, H.; Liu, D.; Wang, C.; Su, B.; and Zheng, C. 2022a. Semmae: Semantic-guided masking for learning masked autoencoders. *Advances in Neural Information Processing Systems*, 35: 14290–14302.

Li, J.; Xu, Q.; He, X.; Liu, Z.; Zhang, D.; Wang, R.; Qu, R.; and Qiu, G. 2025. CFFormer: Cross CNN-Transformer Channel Attention and Spatial Feature Fusion for Improved Segmentation of Low Quality Medical Images. *arXiv preprint arXiv:2501.03629*.

Li, S.; Wu, D.; Wu, F.; Zang, Z.; Li, S.; et al. 2022b. Architecture-agnostic masked image modeling–from vit back to cnn. *arXiv preprint arXiv:2205.13943*.

Li, Y.; Mao, H.; Girshick, R.; and He, K. 2022c. Exploring plain vision transformer backbones for object detection. In *European conference on computer vision*, 280–296. Springer.

Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; and Belongie, S. 2017. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2117–2125.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, 740–755. Springer.

Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; and Berg, A. C. 2016. Ssd: Single shot multibox detector. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, 21–37. Springer.

Liu, Y.; Wu, Y.-H.; Sun, G.; Zhang, L.; Chhatkuli, A.; and Van Gool, L. 2024. Vision transformers with hierarchical attention. *Machine intelligence research*, 21(4): 670–683.

Liu, Z.; Hu, H.; Lin, Y.; Yao, Z.; Xie, Z.; Wei, Y.; Ning, J.; Cao, Y.; Zhang, Z.; Dong, L.; et al. 2022a. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12009–12019.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10012–10022.

Liu, Z.; Mao, H.; Wu, C.-Y.; Feichtenhofer, C.; Darrell, T.; and Xie, S. 2022b. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11976–11986.

Long, J.; Shelhamer, E.; and Darrell, T. 2015. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3431–3440.

Madan, N.; Ristea, N.-C.; Nasrollahi, K.; Moeslund, T. B.; and Ionescu, R. T. 2024. Cl-mae: Curriculum-learned masked autoencoders. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2492–2502.

Mo, S. 2025. The Dynamic Duo of Collaborative Masking and Target for Advanced Masked Autoencoder Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 19484–19492.

Oord, A. v. d.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Ren, S.; Wei, F.; Zhang, S. A. Z.; and Hu, H. 2025. Deepmim: Deep supervision for masked image modeling. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 879–888. IEEE.

Ryali, C.; Hu, Y.-T.; Bolya, D.; Wei, C.; Fan, H.; Huang, P.-Y.; Aggarwal, V.; Chowdhury, A.; Poursaeed, O.; Hoffman, J.; et al. 2023. Hiera: A hierarchical vision transformer without the bells-and-whistles. In *International conference on machine learning*, 29441–29454. PMLR.

Shi, B.; Zhang, X.; Wang, Y.; Li, J.; Dai, W.; Zou, J.; Xiong, H.; and Tian, Q. 2023. Hybrid distillation: Connecting masked autoencoders with contrastive learners. *arXiv preprint arXiv:2306.15876*.

Shi, Y.; Siddharth, N.; Torr, P.; and Kosiorek, A. R. 2022. Adversarial masking for self-supervised learning. In *International Conference on Machine Learning*, 20026–20040. PMLR.

Shin, J.; Lee, I.; Lee, J.; and Lee, J. 2024. Self-Guided Masked Autoencoder. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Tarvainen, A.; and Valpola, H. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30.

Tian, K.; Jiang, Y.; Diao, Q.; Lin, C.; Wang, L.; and Yuan, Z. 2023. Designing bert for convolutional networks: Sparse and hierarchical masked modeling. *arXiv preprint arXiv:2301.03580*.

Van Hasselt, H.; Guez, A.; and Silver, D. 2016. Deep reinforcement learning with double q-learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30.

Wang, L.; Tao, X.; Liu, Q.; and Wu, S. 2024. Rethinking graph masked autoencoders through alignment and uniformity. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 15528–15536.

Wang, W.; Xie, E.; Li, X.; Fan, D.-P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; and Shao, L. 2021. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF international conference on computer vision*, 568–578.

Wu, H.; Xiao, B.; Codella, N.; Liu, M.; Dai, X.; Yuan, L.; and Zhang, L. 2021. Cvt: Introducing convolutions to vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, 22–31.

Xiao, T.; Liu, Y.; Zhou, B.; Jiang, Y.; and Sun, J. 2018. Unified perceptual parsing for scene understanding. In *Proceedings of the European conference on computer vision (ECCV)*, 418–434.

Xie, Z.; Zhang, Z.; Cao, Y.; Lin, Y.; Bao, J.; Yao, Z.; Dai, Q.; and Hu, H. 2022. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9653–9663.

Yuan, L.; Chen, Y.; Wang, T.; Yu, W.; Shi, Y.; Jiang, Z.-H.; Tay, F. E.; Feng, J.; and Yan, S. 2021. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *Proceedings of the IEEE/CVF international conference on computer vision*, 558–567.

Zhang, P.; Dai, X.; Yang, J.; Xiao, B.; Yuan, L.; Zhang, L.; and Gao, J. 2021. Multi-scale vision longformer: A new vision transformer for high-resolution image encoding. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2998–3008.

Zhao, H.; Shi, J.; Qi, X.; Wang, X.; and Jia, J. 2017. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2881–2890.

Zhou, B.; Zhao, H.; Puig, X.; Fidler, S.; Barriuso, A.; and Torralba, A. 2017. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 633–641.

Zhou, J.; Wei, C.; Wang, H.; Shen, W.; Xie, C.; Yuille, A.; and Kong, T. 2022. iBOT: Image BERT Pre-Training with Online Tokenizer. *International Conference on Learning Representations (ICLR)*.

Zhu, R.; Bai, Y.; Yao, T.; Liu, J.; Sun, Z.; Mei, T.; and Chen, C. W. 2024. Teaching Masked Autoencoder With Strong Augmentations. *IEEE Transactions on Neural Networks and Learning Systems*.