

# HiViT: Hierarchical Vision Transformer Meets Masked Image Modeling

Xiaosong Zhang<sup>1\*</sup> Yunjie Tian<sup>1\*</sup> Wei Huang<sup>1</sup> Qixiang Ye<sup>1</sup> Qi Dai<sup>2</sup>

Lingxi Xie<sup>3</sup>

Qi Tian<sup>3</sup>

University of Chinese Academy of Sciences<sup>1</sup>

Fudan University<sup>2</sup>

Huawei Inc.<sup>3</sup>

## Abstract

Recently, masked image modeling (MIM) has offered a new methodology of self-supervised pre-training of vision transformers. A key idea of efficient implementation is to discard the masked image patches (or tokens) throughout the target network (encoder), which requires the encoder to be a plain vision transformer (*e.g.*, ViT), albeit hierarchical vision transformers (*e.g.*, Swin Transformer) have potentially better properties in formulating vision inputs. In this paper, we offer a new design of hierarchical vision transformers named **HiViT** (short for Hierarchical ViT) that enjoys both high efficiency and good performance in MIM. The key is to remove the unnecessary ‘local inter-unit operations’, deriving structurally simple hierarchical vision transformers in which mask-units can be serialized like plain vision transformers. Empirical studies demonstrate the advantageous performance of HiViT in terms of fully-supervised, self-supervised, and transfer learning. In particular, in running MAE on ImageNet-1K, HiViT-B reports a **+0.6% accuracy gain** over ViT-B and a **1.9× speed-up** over Swin-B, and the performance gain generalizes to downstream tasks of detection and segmentation. Code will be made publicly available.

## 1 Introduction

Deep neural networks have been the fundamentals of deep learning [26] and advanced the research fields of computer vision, natural language processing, *etc.*, in the past decade. Recently, the computer vision community has witnessed the emerge of vision transformers [14, 30, 47, 60, 10, 27], transplanted from the language models [46, 12], that replaced the dominance of convolutional neural networks [25, 23, 39]. They have the ability of formulating long-range feature dependencies, which naturally benefits visual recognition especially when long-range relationship is important.

There are mainly two families of vision transformers, namely, the plain vision transformers [14, 43] and the hierarchical vision transformers [30, 47, 13, 4], differing from each other in whether multi-resolution feature maps are used. While the latter is believed to capture the nature of vision signals (most convolution-based models have used the hierarchical configuration), but used some spatial local operations (*i.e.*, early-stage self-attentions with shifting window). These models can encounter difficulties when the tokens need to be flexibly manipulated. A typical example lies in masked image modeling (MIM), a recent methodology of pre-training vision transformers [1, 20, 54] – a random part of image patches are hidden from input, and it is difficult for the hierarchical models to

\*Equal Contribution.

determine whether each pair of tokens need to communicate, unlike the plain models. Essentially, this is because hierarchical vision transformers have used non-global operations (*e.g.*, window attentions) between the masking units<sup>2</sup>. Hence, unlike the plain vision transformers that can serialize all tokens for acceleration, the hierarchical vision transformers must maintain the two-dimensional structure, keeping the dummy (masked) tokens throughout the encoder. Consequently, as shown in [54], the training speed of hierarchical transformers is  $2\times$  slower than that using plain transformers, and very few works chose to follow this direction.

In this paper, we start with categorizing the operations in hierarchical vision transformers into ‘intra-unit operations’, ‘global inter-unit operations’, and ‘local inter-unit operations’. We note that plain vision transformers only contain ‘intra-unit operations’ (*i.e.*, patch embedding, layer normalization, MLP) and ‘global intra-unit operations’ (*i.e.*, global self-attentions), hence the units’ spatial coordinates can be discarded and the units can be serialized for efficient computation, like in MAE [20]. That said, for hierarchical vision transformers, it is the ‘local inter-unit operations’ (*i.e.*, shifting-window self-attentions, patch merging) that calls for extra judgment based on the units’ spatial coordinates and obstructs the serialization as well as removing the masked units.

A key observation of this paper lies in that ‘local inter-unit operations’ do not contribute much for recognition performance – what really makes sense is the hierarchical design (*i.e.*, multi-scale feature maps) itself. Hence, to fit hierarchical vision transformers to MIM, we remove the ‘local inter-unit operations’, resulting in a simple hierarchical vision transformer that absorbs both the flexibility of ViT [14] and the superiority of Swin Transformer [30]. There are usually 4 stages of different resolutions in hierarchical vision transformers where the 3rd stage has the largest number of layers and we call it the main stage. We remove the last stage of Swin and switch off all local inter-unit window attentions, only keeping the global attention between tokens in the main stage. In practice, the last stage is merged into the main stage (to keep the model FLOPs unchanged) and local window attentions in the early-stage are replaced by an intra-unit multi-layer perceptron with same FLOPs. With these minimal modifications, we remove all redundant ‘local inter-unit operations’ in hierarchical vision transformers, where only the simplest hierarchical structure is adopted. Compared to the plain ViTs, our model only adds only several spatial merge operations and MLP layers before the main stage. The resulting architecture is named **HiViT** (short for Hierarchical ViT), which has the ability of modeling hierarchical visual signals yet all tokens are maximally individual and remain flexibility for manipulation. Meanwhile, our HiViT maintains the ViT paradigm, which is very simple to implement compared to other hierarchical vision transformers.

We perform fully-supervised classification experiments on ImageNet-1K to validate the superiority of HiViT. Lying between ViT and Swin Transformer, HiViT enjoys consistent accuracy gains over both the competitors, *e.g.*, HiViT-B reports a 83.8% top-1 accuracy, which is  $+2.0\%$  over ViT-B and  $+0.3\%$  over Swin-B. With extensive ablation studies, we find that removing the ‘local inter-unit operations’ does not harm the recognition performance, yet the hierarchical structure and relative positional encoding (not ‘local inter-unit operations’) slightly but consistently improves the classification accuracy. This makes HiViT applicable to a wide range of visual recognition scenarios.

Continuing to MIM, the advantages of HiViT become clearer. With 800 epochs of MIM-based pre-training and 100 epochs of fine-tuning, HiViT-B reports 84.2% top-1 accuracy on ImageNet-1K, which is  $+0.6\%$  over ViT-B (using MAE [20], pre-training for 1600 epochs) and  $+0.2\%$  over Swin-B (using SimMIM [54]). More importantly, HiViT enjoys the efficient implementation that discards all masked patches (or tokens) at the input stage, and hence the training speed is  $1.9\times$  as fast as that of Sim-

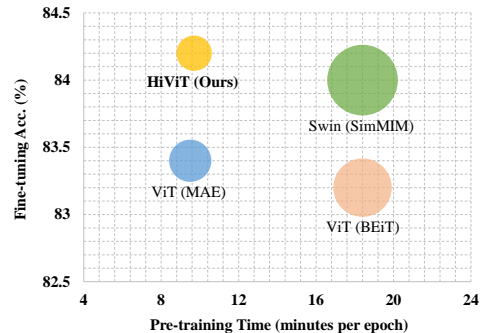


Figure 1: Self-supervised pre-training of HiViT is significantly faster than Swin with SimMIM [54] and the result is better than ViT trained with MAE [20] and BEiT [1]. Circle size denotes memory requirement. All the models are in base scale.

<sup>2</sup>The minimum size of the masked pixels when executing MIM is defined as masking unit. For example, when the input image size is  $224 \times 224$ , the masking unit size for MAE [20] is  $16 \times 16$ .

MIM, since the original Swin Transformer must forward-propagate the full token set (Fig. 1). The advantages persist to other visual recognition tasks, including linear probing (71.3% top-1 accuracy) on ImageNet-1K, semantic segmentation (48.3% mIoU) on the ADE20K dataset [59], and object detection (49.5% AP) and instance segmentation (43.8% AP) on the COCO dataset [28] ( $1\times$  training schedule). These results validate that removing ‘local inter-unit operations’ does not harm generic visual recognition.

The core contribution of this paper is HiViT, a hierarchical vision transformer architecture that is off-the-shelf for a wide range of vision applications. In particular, with masked image modeling being a popular self-supervised learning paradigm, HiViT has the potential of being directly plugged into many existing algorithms to improve their effectiveness and efficiency in learning visual representations from large-scale, unlabeled data.

## 2 Related Work

### 2.1 Vision Transformers

Vision transformers [14] were adapted from the natural language processing (NLP) transformers [46, 12], opening a new direction of designing visual recognition models with weak induction bias [42]. Early vision transformers [14] mainly adopted the plain configuration, and efficient training methods are strongly required [43]. To cater for vision-friendly priors, Swin Transformer [30] proposed a hierarchical architecture that contains multi-level feature maps and validated good performance in many vision problems. Since then, various efforts emerged in improving hierarchical vision transformers, including borrowing design experiences from convolutional neural networks [47, 51, 45], adjusting the design of self-attention geometry [13, 55], designing hybrid architectures to integrate convolution and transformer modules [38, 16, 10, 35], *etc.*

Essentially, there is a tradeoff between plain and hierarchical vision transformers – in terms of whether strong induction bias is to be introduced. As we shall see later, the increase of induction bias may weaken the flexibility and thus efficiency of applying vision transformers to particular scenarios (*e.g.*, masked image modeling). In this paper, we design a hierarchical vision transformer that maximally discards induction bias, achieving both high efficiency and good performance.

### 2.2 Self-Supervised Learning and Masked Image Modeling

In the context of computer vision, self-supervised learning aims to learn compact visual representations from unlabeled data. The key to this goal is to design a pretext task that sets a natural constraint for the target model to achieve by tuning its weights. The existing pretext tasks are roughly partitioned into three categories, namely, **geometry-based** proxies that were built upon the spatial relationship of image contents [50, 33, 17], **contrast-based** proxies that assumed that different views of an image shall produce related visual features [21, 6, 18, 3, 2, 53, 41], and **generation-based** proxies that required visual representations to be capable of recovering the original image contents [58, 34, 20, 1, 40]. After the self-supervised learning (*a.k.a.* pre-training) stage, the target model is often evaluated by fine-tuning in a few downstream recognition tasks – the popular examples include image classification at ImageNet-1K [11], semantic segmentation at ADE20K [59], object detection and instance segmentation at COCO [29], *etc.*

We are interested in a particular generation-based method named masked image modeling (MIM) [1, 20]. The flowchart is straightforward: some image patches (corresponding to tokens) are discarded, the target model receives the incomplete input and the goal is to recover the original image contents. MIM is strongly related to the masked language modeling (MLM) task in NLP. BEiT [1] transferred the task to the computer vision community by masking the image patches and recovering the tokens produced by a pre-trained model (known as the tokenizer). MAE [20] improved the MIM framework by only taking the visible tokens as input and computing loss at the pixel level – the former change largely accelerated the training procedure as the computational costs of the encoder went down. The follow-up works explored different recovery targets [49], more complicated model designs [15, 6], and other pretext tasks.

It is worth noting that MIM matches plain vision transformers very well because each token is an individual unit and only the unmasked tokens are necessary during the pre-training process. The properties does not hold for hierarchical vision transformers, making them difficult to inherit the good

properties (*e.g.*, training efficiency). Although SimMIM [54] tried to combine Swin Transformer with MIM, it uses all tokens, including those corresponding to the masked patches, shall be preserved during the encoder stage, incurring much heavier computational costs. In this paper, we present a hierarchical vision transformer that is free of such burden.

### 3 Hierarchical Vision Transformer for Masked Image Modeling

#### 3.1 Preliminaries

Masked image modeling (MIM) is an emerging paradigm of self-supervised visual representation learning. The flowchart involves feeding a partially masked image to the target model and training the model to recover it. Mathematically, let the target model be  $f(\mathbf{x}; \boldsymbol{\theta})$  where  $\boldsymbol{\theta}$  denotes the learnable parameters. Given a training image,  $\mathbf{x}$ , it is first partitioned into a few patches,  $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_M\}$ , where  $M$  is the number of patches. Then, MIM randomly chooses a subset  $\mathcal{M}' \subset \{1, 2, \dots, M\}$ , feeds the patches with IDs in  $\mathcal{M}'$  (denoted as  $\mathcal{X}'$ ) into the target model  $f(\mathbf{x}; \boldsymbol{\theta})$  (*a.k.a.*, the encoder), and appends a decoder to it, aiming at recovering the original image contents, either tokenized features [1] or pixels [20], at the end of the decoder. If  $f(\mathbf{x}; \boldsymbol{\theta})$  is able to solve the problem, it is believed that the parameters have been well trained to extract compact visual features.

An efficient vision model that fits MIM is the vanilla vision transformer, abbreviated as ViT [14]. In ViT, each image patch is transferred into a token (*i.e.*, a feature vector), and the tokens are propagated through a few transformer blocks for visual feature extraction. Let there be  $L$  blocks, the  $l$ -th block takes the token set of  $\mathcal{U}^{(l-1)}$  as input and outputs  $\mathcal{U}^{(l)}$ , and  $\mathcal{U}^{(0)} \equiv \mathcal{X}$ . The main part of each block is self-attention, for which three intermediate features are computed upon  $\mathbf{u}_m^{(l-1)}$ , namely the query, key, and value, denoted as  $\mathbf{q}_m^{(l-1)}$ ,  $\mathbf{k}_m^{(l-1)}$ , and  $\mathbf{v}_m^{(l-1)}$ , respectively. Based on these quantities, the self-attention of  $\mathbf{z}_m^{(l-1)}$  is computed by  $\text{SA}(\mathbf{z}_m^{(l-1)}) = \text{softmax}\left[\frac{\mathbf{q}_m^{(l-1)} \cdot \mathbf{k}_1^{(l-1)\top}, \dots, \mathbf{q}_m^{(l-1)} \cdot \mathbf{k}_M^{(l-1)\top}}{\sqrt{D_{\text{key}}}}\right] \cdot \frac{1}{\sqrt{D_{\text{key}}}} \left[\mathbf{v}_1^{(l-1)}, \dots, \mathbf{v}_M^{(l-1)}\right]^\top$ , where  $1/\sqrt{D_{\text{key}}}$  is a scaling vector. Auxiliary operations, including layer normalization, multi-layer perceptron, skip-layer connection, are applied after the self-attention computation. ViT has been applied to a series of vision problems, but we emphasize its particular efficiency on MIM, which lies in that the tokens not in  $\mathcal{X}'$  can be discarded at the beginning of encoder, decreasing the complexities of the pre-training process by a factor of  $M/|\mathcal{M}'|$  (*e.g.*, 4 in the regular setting of MAE [20]).

Intuitively, hierarchical vision transformers (*e.g.*, Swin Transformer) are better at capturing multi-level visual features. It has three major differences from ViT: (i) the architecture is partitioned into a few stages and the spatial resolution, rather than being fixed, is gradually shrunk throughout the forward propagation; (ii) to handle relatively large token maps, the self-attention computation is constrained within a grid of windows, and the window partition is shifted across layers; (iii) global positional encoding is replaced by relative positional encoding – this is to fit the window attention mechanism. Although hierarchical vision transformers report higher visual recognition accuracy, these models are not so efficient as ViT in terms of MIM, and the reasons are revealed in the next part. Consequently, few prior works have tried the combination – as an example, SimMIM [54] fed the entire image (the masked patches are replaced with mask tokens which are learnable) into the encoder, resulting in heavier computational costs in time and memory.

#### 3.2 HiViT: Efficient Hierarchical Transformer for MIM

We pursue for the efficient implementation of MAE [20], *i.e.*, only the active (unmasked) tokens are fed into the encoder – mathematically, we are always dealing with a squeezed list of  $|\mathcal{M}'|$  tokens. The major difficulty of integrating it with hierarchical vision transformers (*e.g.*, Swin Transformers) lies in the ‘local inter-unit operations’, which make it difficult to serialize the tokens and abandon the inactive (masked) ones. To remove them, we first set the masking unit size to be the token size at the main stage – for Swin Transformers, this is the 3rd stage that is the major part (*e.g.*, for Swin-B, the 3rd stage has 18 blocks, and the entire architecture has 24 blocks). The masking unit is  $16 \times 16$  pixels that aligns with the constant token size of ViT. Then, we adjust the model as follows:

- For the operations after the main stage, we do not allow the patch merging that mixes active and inactive patches. For simplicity, we directly remove the last (4th) stage in the Swin

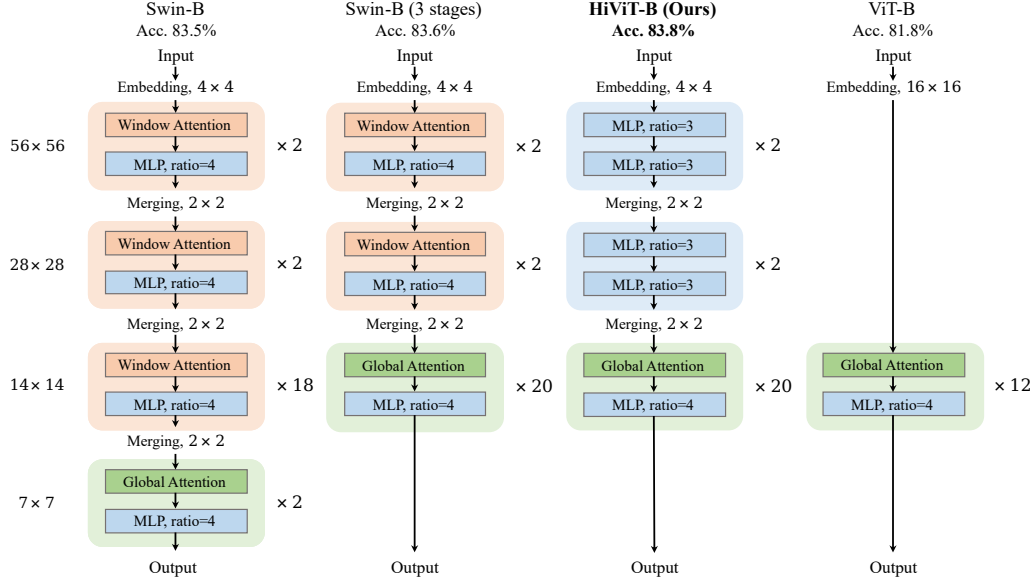


Figure 2: Comparison of the architectures of Swin Transformer, ViT, and the proposed HiViT.

Table 1: Configurations for HiViT variants.

Model	Depth			Dim			Heads			Params (M)	FLOPs (G)
	$56^2$	$28^2$	$14^2$	$56^2$	$28^2$	$14^2$	$56^2$	$28^2$	$14^2$		
HiViT-T (Tiny)	1	1	10	96	192	384	-	-	6	19.2	4.6
HiViT-S (Small)	2	2	20	96	192	384	-	-	6	37.5	9.1
HiViT-B (Base)	2	2	20	128	256	512	-	-	8	66.4	15.9

Transformer, which has only two blocks, and append the same number of blocks to the 3rd stage. Since the 3rd stage has a smaller token dimensionality, such an operation saves both trainable parameters, while changing the ‘local inter-unit operations’ in the 3rd stage to ‘global inter-unit operations’. As a result, the model’s recognition performance becomes better (Fig. 2).

- For the operations prior to the main stage, we do not allow the window attention in the former 2 stages. That said, we remove the shift window of Swin and do not introduce any other ‘local inter-unit operations’ such as window attentions or convolutions. As an alternative, we only employ MLP block (replacing the self-attention by another mlp layer) for the former 2 stages. Surprisingly, as demonstrated in Fig. 2, this modification brings 0.2% performance improvement without bells and whistles. Compared to plain ViT as shown in Fig. 2, the derived architecture possesses hierarchical property and only requires two MLP blocks but enjoys much better performance on both self-/fully-supervised learning.

The above procedure produces an architecture between Swin Transformer (hierarchical) and ViT (plain). We illustrate the procedure in Figure 2. The resulting architecture, named HiViT (short for Hierarchical ViT), **is structurally simple and brings an efficient implementation for MIM**. Specifically, HiViT abandons all the ‘local inter-unit operations’ in the entire architecture, therefore the masked image patches can be discarded at the input layer and its computation can be eliminated in all stages. As a result, HiViT enjoys both the effectiveness of hierarchical vision transformers to capture visual representations (*i.e.*, the recognition accuracy is much higher than ViT) and the efficiency of plain vision transformers in the masked image modeling task (*i.e.*, the efficient implementation of MAE [20] can be directly transplanted, making HiViT almost  $2\times$  faster than Swin Transformer in MIM). Detailed results are provided in the experimental part.

Table 2: Comparison of image classification on ImageNet-1K for different models. All models are trained and evaluated with  $224 \times 224$  resolution on ImageNet-1K in default, unless otherwise noted.

Model	Params (M)	FLOPs (G)	Top-1 (%)	Model	Params (M)	FLOPs (G)	Top-1 (%)
DeiT-S/16 [43]	22.1	4.5	79.8	ResNet-152 [23]	60.0	11.0	78.3
PVT-S[48]	24.5	3.8	79.8	PVT-L [47]	61.4	9.8	81.7
Swin-T [31]	28.3	4.5	81.2	DeiT-B/16 [43]	86.7	17.4	81.8
CvT-13 [51]	20.0	4.5	81.6	CrossViT-B [4]	104.7	21.2	82.2
CaiT-XS-24 [44]	26.6	5.4	81.8	T2T-ViT-24 [56]	64.1	14.1	82.3
HiViT-T (Ours)	19.2	4.6	<b>82.1</b>	CPVT-B [9]	88.0	17.6	82.3
CvT-21 [51]	32.0	7.1	82.5	TNT-B [19]	65.6	14.1	82.8
UFO-ViT-M [37]	37.0	7.0	82.8	ViL-B [57]	55.7	13.4	83.2
Swin-S [31]	49.6	8.7	83.1	UFO-ViT-B [37]	64.0	11.9	83.3
ViL-M [57]	39.7	9.1	83.3	CaiT-M24 [44]	185.9	36.0	83.4
CaiT-S36 [44]	68.0	13.9	83.3	Swin-B [31]	87.8	15.4	83.5
HiViT-S (Ours)	37.5	9.1	<b>83.5</b>	HiViT-B (Ours)	66.4	15.9	<b>83.8</b>

## 4 Experiments

We first conduct fully supervised experiments with labels using the proposed HiViT on ImageNet dataset [11]. Then, HiViT models are tested using masked image modeling self-supervised methods (MIM) [20]. The self-supervised pre-trained models are also transferred to downstream tasks including object detection on COCO dataset [29] and semantic segmentation on ADE20K [59]. We provide the ablation studies about our methods on both fully-/self-supervised learning.

### 4.1 ImageNet Classification with Labels

**Training Settings** We first evaluate our models with fully supervised learning on ImageNet-1K [11] which contains 1.28M training images and 50K validation ones divided into 1,000 categories. We follow Swin [30] and use the same training settings without other tricks. Specifically, we use AdamW optimizer [32] with an initial learning rate of 0.001, a weight decay of 0.05, batch size of 1024, a cosine decay learning rate scheduler, and a linearly warm-up for 20 epochs. All the models are trained for 300 epochs with augmentation and regularization strategies of [30] and exponential moving average (EMA) technique. The input size is  $224 \times 224$  in default. The output feature of 3rd stage is followed by an average pooling layer and then a classifier layer. We adopt drop path rate of 0.05, 0.3, and 0.5 for HiViT-T/S/B.

**Model Configurations** Three models including HiViT-T/S/B are tested on fully supervised learning and their configurations are shown in Tab. 1. The “Depth” represents the block number on different stages ( $56^2$ ,  $28^2$ , and  $14^2$  are the 1st, 2nd, and 3rd stage respectively). The “Dim” and “Heads” represent the dimension and attention head of 3 stages. We align our models on FLOPs, and the parameters are less compared to other works. We report the inference throughput speed in Tab. 5 by testing the  $224^2$  images on V100 GPU with the same script.

**ImageNet Results** The fully supervised training results are shown in Tab. 8. Compared to vanilla ViT models, all the HiViT models report dominant results. HiViT-T/B surpass DeiT-S/B models by 2.3% and 2.0% respectively with similar flops and fewer parameters. Compared to its follow-ups, our models still show competitive results. In particular, HiViT-T/S/B beat Swin-T/S/B by 0.9%, 0.4%, and 0.3% respectively with similar complexities and fewer parameters. All the results do not introduce any other tricks compared to Swin [30]. In addition, all our models are parameter friendly. For example, compared to Swin-T/S/B models, HiViT-T/S/B enjoy 32.2%, 24.4% and 24.4% fewer parameters. We note that our models are structurally simple which offer a promising baseline for future research.

**Ablation Studies** We conduct fully supervised ablation studies to show the advantages of our method. In this part, we inherit the training settings above and the results are shown in Tab. 3.

Table 3: Ablations of fully-supervised training on ImageNet-1K.

Model	Setting	Dim	Depth				RPE	Win. Att.	Params (M)	FLOPs (G)	Top-1 (%)
			56 <sup>2</sup>	28 <sup>2</sup>	14 <sup>2</sup>	7 <sup>2</sup>					
Swin-B	-	512	2	2	18	2	✓	✓	88.0	15.4	83.5
HiViT-B	Stage4	512	2	2	20	-	✓	✓	66.3	16.0	83.6
	Win. Att.	512	2	2	20	-	✓	✗	66.4	15.9	<b>83.8</b>
	RPE	512	2	2	20	-	✗	✗	66.3	15.9	83.5
ViT-B	Hierarchical	512	-	-	24	-	✗	✗	76.7	15.8	82.9
	Deep	768	-	-	12	-	✗	✗	86.6	17.5	81.8

The ‘Setting’ represents that the modules we remove from top to bottom. The ‘Dim’ represents the dimension on 14<sup>2</sup> resolution. The ‘Depth’ represents the block numbers of the corresponding stages. The ‘RPE’ is the relative position embedding and the “Win. Att.” denotes whether there are window attentions in the model. As we can see in the Tab. 3, removing the last stage (Stage4) from Swin-B (using global attention for stage3 simultaneously) brings a performance improvement of 0.1% which implies that the last stage is unnecessary. Replacing window attention with MLP blocks in the former 2 stages (Win. Att.) boosts the performance to 83.8%, which demonstrates that window attention is unnecessary in early stages. The RPE is important and getting rid of that (RPE) will harm the performance about 0.3%. If we abandon the former 2 stages and down-sample 16× using the patch embedding like plain ViT but increasing the block number to 24 (Hierarchical), the performance will decrease from 83.5% to 82.9%. However, that is still higher than 81.8% of plain ViT (Deep), which implies that hierarchical input module is important and a deeper architecture is much better than a shallow one.

## 4.2 Self-Supervised Learning Results

**Experimental Details** For self-supervised pre-training, we use ImageNet-1K training dataset without using labels, and test the pre-trained models by fine-tuning and linear probing metrics in validation dataset. We inherit the pre-training settings of MAE [20] for pre-training. Specifically, we set the mask ratio to 75% in default. The normalized target trick is also adopted. We use the AdamW optimizer [32] with the an initial learning rate of  $1.5 \times 10^{-4}$ , a weight decay of 0.05, and the learning rate follows the cosine decay learning schedule with a warm-up for 40 epochs. The batch size is set to 4096 and the input size is  $224 \times 224$ . The overall pipeline is an encoder-decoder framework and the decoder is designed to have 6 transformer layers followed by a reshape operation to cast the feature to  $3 \times 224 \times 224$ . As for data augmentation, we only employ random cropping and random horizontal flip. We test HiViT-B model in this part and the model is pre-trained for 300 and 800 epochs, which are then evaluated using fine-tuning and linear-probing metrics. As for fine-tuning, we inherit the training settings from [20] and all the models are trained for 100 epochs using AdamW optimizer with a warm-up for 5 epochs, a weight decay of 0.05, and input size of  $224 \times 224$ . We use the layer-wise learning rate decay of 0.65. The initial learning rate is set to  $5 \times 10^{-4}$  and batch size is set to 1024. As for linear probing, we train all the models for 100 epochs using LARS [24] optimizer with the batch size of 16,384 and learning rate of 0.1.

**Fine-tuning** The fine-tuning and linear probing results are provided in Tab. 4 and only the encoder part is used to test. As shown in the Tab. 4, the HiViT-B (300e) version achieves 0.4% and 0.2% performance improvement than MAE models which are pre-trained for 1,600 epochs. Our longer training schedule version (800e) attains the dominant result of 84.2%, which outperforms MAE (1600e) by 0.6% and SimMIM (Swin) by 0.2%. Compared to other methods including CAE, BEiT and iBOT, etc, HiViT-B model also shows superior results: +1.0% for BEiT, +0.6% for CAE, and +0.2% for MaskFeat.

**Linear Probing** We evaluated the pre-trained models using linear probing metric, where all the parameters of the encoder are frozen except for a learnable classifier layer. From Tab. 4, we can see that HiViT-B model achieves good result of 71.3%, which is the best performance compared to all the MIM based methods. For example, the 800e model surpasses MAE (1600e) by 3.3% with fewer

Table 4: Self-supervised learning results. The fine-tuning and linear probing results.

Method	Network	Params	Supervision	Encoder	Epochs	FT (%)	LIN (%)
BEiT [1]	ViT-B	86	DALLE	100%	400	83.2	-
CAE [7]	ViT-B	86	DALLE	100%	800	83.6	68.3
MaskFeat [49]	ViT-B	86	HOG	100%	800	84.0	-
SimMIM [54]	ViT-B	86	Pixel	100%	800	83.8	68.7
SimMIM [54]	Swin-B	86	Pixel	100%	800	84.0	-
MAE [20]	ViT-B	86	Pixel	25%	1600	83.6	68.0
Ours	HiViT-B	66	Pixel	25%	300	83.8	-
Ours	HiViT-B	66	Pixel	25%	800	<b>84.2</b>	<b>71.3</b>

pre-training epochs and the same training settings. The result also outperforms SimMIM (ViT-B) by 2.6% and CAE (ViT-B) by 2.8%.

**Training Efficiency** HiViT only requires the active tokens as inputs so that our method enjoys the efficiency during the MIM pre-training. As shown in Tab. 5, we report the pre-training speed of MAE (ViT-B), SimMIM (Swin-B), and our HiViT-B with different input sizes. All the results represent the pre-training time (minutes) of 1 epoch on  $8 \times V100$  GPUs. As the input image is  $192 \times 192$ , HiViT-B only takes 7.4 minutes per epoch, which is faster about  $1.9 \times$  than SimMIM and comparable with MAE. HiViT-B takes about 9.7 minutes when the input is  $224 \times 224$ , which is  $1.9 \times$  faster than SimMIM and comparable with MAE. We note that we use 6 decoder blocks with 512 dimension in default. Decreasing the decoder block number or the dimension also accelerate the pre-training speed without affecting our results. For example, setting the decoder block number to 4 and dimension to 384 only requires about 8 minutes per epoch and achieves 83.8% performance.

Table 5: Pre-training efficiency comparison with different input sizes.

Input Size	ViT-B (MAE)	Swin-B (SimMIM)	HiViT-B (Ours)
$192 \times 192$	7.2	14.2	7.4
$224 \times 224$	9.5	18.4	9.7

**Ablation Studies** We perform some experiments to ablate our methods (see Tab. 6) and all the results are attained by pre-training for 300 epochs. As shown, we test the self-supervised learning pre-training by setting different block number for stage 1, 2, and 3 (referred as to  $56^2$ ,  $28^2$ , and  $14^2$  respectively). The default setting (#0) achieves 83.8% performance with 2–2–20 block setting. Decreasing the block number for stage 1, 2 and increasing for stage 3 (#1) brings more parameters and better performance of 83.9%, which is comparable to SimMIM result (84.0%) by pre-training for 800 epochs with Swin-B. We note that the 71.9M parameters are still much lower than 87.8M of Swin-B. Removing the stage-1 (#3) or stage-2 (#2) both harm the performance to 83.6% and 83.7% respectively, which demonstrates that the hierarchical architecture before the main stage is important and brings performance improvement. In addition, the result of #3 is lower than #2 denotes that the first stage seems more important than the second one. Removing both the former 2 stages attains a result of 83.6%, which further verifies the importance of the hierarchical architecture.

Table 6: Ablations by pre-training for 300 epochs and fine-tuning for 100 epochs (FT). The  $56^2$ ,  $28^2$ , and  $14^2$  denote the 1st, 2nd and 3rd stage respectively. The ID #0 is the default setting.

ID	Depth			Params (M)	FLOPs (G)	FT (%)
	$56^2$	$28^2$	$14^2$			
#0	2	2	20	66.4	15.9	83.8
#1	1	1	22	71.8	15.9	<b>83.9</b>
#2	2	0	22	71.1	15.9	83.7
#3	0	2	22	72.3	15.9	83.6
#4	0	0	24	77.1	16.0	83.6

### 4.3 Transfer to Dense Prediction Tasks

**Experimental Details** We transfer the self-supervised pre-trained models above to object detection on COCO and semantic segmentation on ADE20K. We follow the convention to perform object detection and semantic segmentation experiments. For the COCO experiments, we use the Mask R-CNN [22] head implemented by MMDetection library [5]. We use the AdamW optimizer [32]



Table 7: Downstream task fine-tuning results transferred from self-supervised pre-training.

Method	Network	Params	Pre-train data	COCO		ADE20K mIoU
				AP <sup>box</sup>	AP <sup>mask</sup>	
Supervised [20]	ViT-B	86	IN1K w/ labels	47.9	42.9	47.0
MoCo v3 [8]	ViT-B	86	IN1K	45.5	40.5	47.3
BEiT [1]	ViT-B	86	IN1K+DALLE	42.1	37.8	47.1
CAE [7]	ViT-B	86	IN1K+DALLE	49.2	43.3	<b>48.8</b>
MAE [20] (1600e)	ViT-B	86	IN1K	48.4	42.6	48.1
Ours (800e)	HiViT-B	66	IN1K	<b>49.5</b>	<b>43.8</b>	48.3

with an initial learning rate of  $3 \times 10^{-4}$  which decays by  $10\times$  after the 9-th and 11-th epochs. The layer-wise decay rate is set to 0.75 and  $1\times$  training schedule (12 epochs) is adopted. We also apply multi-scale training strategy and single-scale testing. For ADE20K, we use the UperNet [52] head following BEiT [1]. We also choose AdamW optimizer and the learning rate is  $4 \times 10^{-4}$ . We totally train the model for 160 iterations and the batch size is 16. The input resolution is  $512 \times 512$  without using multi-scale testing.

**Objection Detection on COCO.** We transfer the same settings of CAE [7] to test our model in MS-COCO. We choose the 5-, 9-, 13-, 19-th blocks as inputs for later FPN network. As shown in Tab. 7, we compare the performance with the state-of-the-art methods. Compared to BEiT [1] (we borrow the results from CAE [7]), HiViT-B shows superior results over 7.4% AP<sup>box</sup> and 6.0% AP<sup>mask</sup> respectively. MAE [20] (1600e) achieves the 48.4% AP<sup>box</sup> and 42.6% AP<sup>mask</sup> results, which is lower than our 49.5% and 43.8% respectively. CAE [7] improves the performances to 49.2% AP<sup>box</sup> and 43.3% AP<sup>mask</sup> by pre-training for 800 epochs. But the results still are below than our results by 0.3% and 0.8% respectively even though CAE [7] uses image tokenizer described in DALLE [36].

**Semantic Segmentation on ADE20K.** The results on ADE20K are shown in Tab. 7. We note that we do not introduce any other tricks and test all the models using the same settings. We report the mean intersection over union (mIoU) performances. As shown, MoCo-v3 reports the 47.3% mIoU result by pre-training for 300 epochs, which is lower than our 48.3%. BEiT [1], CAE[7], and MAE [20] report the performance of 47.1%, 48.8%, and 48.1% respectively. By pre-training for 1600 epochs, MAE achieves the 48.1% mIoU. Compared to these state-of-the-art methods, HiViT-B, by pre-training for 800 epochs, reports the 48.3% result, which is higher than all the methods in apart from CAE, which uses tokenizer of DALLE [36].

## 5 Conclusions

This paper presents a hierarchical vision transformer named HiViT. Starting with Swin Transformers, we remove redundant operations that cross the border of tokens in the main stage, and show that such modifications do not harm, but slightly improve the model’s performance in both fully-supervised and self-supervised visual representation learning. HiViT shows a clear advantage in integrating with masked image modeling, on which the efficient implementation on ViT can be directly transplanted, accelerating the training speed by almost 100%. We expect that HiViT becomes an off-the-shelf replacement of ViT and Swin Transformers in the future research.

**Limitations** Despite the improvement observed in the experiments, our method has some limitations. The most important one lies in that the masking unit size is fixed – this implies that we need to choose a single ‘main stage’. Fortunately, the 3rd stage of Swin Transformers contribute most parameters and computations, hence it is naturally chosen, however, the method may encounter difficulties in the scenarios that no dominant stages exist. In addition, we look forward to more flexible architecture designs that go beyond the constraints – a possible solution lies in modifying low-level code (e.g., CUDA) to support arbitrary and variable grouping of tokens.

**Societal Impacts** Our research focus on (i) designing efficient architectures for deep neural networks and (ii) self-supervised learning. These two topics have been widely studied in the community, and our work does not have further societal impacts than others.

## Appendix

### A Architecture Comparison of Fully-Supervised Learning on ImageNet-1K

We compare different transformer architectures including **plain transformers**, **hierarchical transformers**, and **hybrid ones** on fully-supervised learning using ImageNet-1K and the results are shown in Tab. 8. We can see that plain transformers usually do not contain local inter-unit operations, which implies that these models are structurally simple generally. However, these models suffer poor performance than hierarchical ones. To capture hierarchical property for transformers, researchers introduce complex local inter-unit operations (such as spatial-reduction attention and window attention, etc., as shown in the table) to plain transformer, so that the models cater to visual priors and attain better results. Our HiViT, as shown, enjoys the hierarchical attributes without local inter-unit operations, which makes our model structurally simple and MIM friendly. In addition, HiViT gets better results without bells and whistles. The hybrid transformers bring convolutional layers (except for patch embedding) to transformer and usually attain the dominant results than the former two. But these models are unfriendly to MIM self-supervised task because of the sliding prior of convolutional layers. Moreover, our self-supervised HiViT-B achieves competitive result of 84.2%.

Table 8: ImageNet-1K results of different transformer architectures.

Model type	Model	Local inter-unit operations	Params (M)	FLOPs (G)	Top-1 (%)
Plain Transformers	ViT-B/16 [43]	-	86.7	17.4	81.8
	CrossViT-B [4]	-	104.7	21.2	82.2
	CaiT-M24 [44]	-	185.9	36.0	83.4
	T2T-ViT-24 [56]	-	64.1	14.1	82.3
	TNT-B [19]	-	65.6	14.1	82.8
Hierarchical Transformers	PVT-L [47]	spatial-reduction attention	61.4	9.8	81.7
	ViL-B [57]	window attention	55.7	13.4	83.2
	Swin-B [31]	shifted window attention	87.8	15.4	83.5
	HiViT-B (Ours)	-	66.4	15.9	<b>83.8</b>
Convolution + Transformers	Conformer[35]	CNN branch	83.3	23.3	84.1
	CoAtNet-2[10]	depth-wise convolution	75.0	15.7	84.1
	CSwin-B[13]	$\left\{ \begin{array}{l} \text{convolution in LePE} \\ \text{cross-shaped window attention} \end{array} \right.$	78.0	15.0	84.2
Self-supervised	HiViT-B (Ours)	-	66.4	15.9	<b>84.2</b>

### B Improved Results on COCO and ADE20K

In dense prediction tasks such as object detection and semantic segmentation, the feature pyramid is a key component. Our experiments in Tab. 7 use the same settings as the ViT model using self-supervised pre-training in [20, 7]. This setting extracts intermediate features and up-samples/down-samples them by deconvolution/convolution for pyramid feature generation, which works for plain transformers but does not take advantage of our hierarchical structure.

In order to fully exploit the capabilities of our hierarchical vision transformer, we perform an improved experimental setting following [31]. We use the three resolution features (with strides of 4, 8, 16) generated by stage-1/-2/-3, and add a stride-32 feature which is down-sampled from the stage-3 block to align with the standard feature pyramid generation. With this change, HiViT achieves improved results on COCO and ADE20K, as shown in Tab. 9. In particular, it achieves 51.2% AP<sup>box</sup>, 44.2% AP<sup>mask</sup> in COCO and 51.2% mIoU in ADE20K.

Table 9: Improved downstream task fine-tuning results on COCO and ADE20K.

Method	Network	Params	Pre-train data	COCO		ADE20K mIoU
				AP <sup>box</sup>	AP <sup>mask</sup>	
Supervised [20]	ViT-B	86	IN1K w/ labels	47.9	42.9	47.0
MoCo v3 [8]	ViT-B	86	IN1K	45.5	40.5	47.3
BEiT [1]	ViT-B	86	IN1K+DALLE	42.1	37.8	47.1
CAE [7]	ViT-B	86	IN1K+DALLE	49.2	43.3	48.8
MAE [20] (1600e)	ViT-B	86	IN1K	48.4	42.6	48.1
Ours	HiViT-B	66	IN1K	49.5	43.8	48.3
Ours (improved)	HiViT-B	66	IN1K	51.2	44.2	51.2

## C The Processing Pipeline of HiViT in MIM Pre-training

In MIM pre-training, HiViT uses the encoder to process the visible tokens of the input image as shown in Fig. 3. Here, since we have multi-resolution image patches, we call the smallest unit that may be masked as a mask-unit. For an input image, we first align the patch embedding to get a feature vector of shape  $M \times 4 \times 4 \times 128$ , where  $M$  is the total number of all mask-units (e.g.,  $M = 196$  according to the common practice). Then, we randomly select a certain proportion (25%) of mask-units to obtain a feature vector of  $M' \times 4 \times 4 \times 128$ , where  $M'$  is the number of visible mask-units which is often much smaller than  $M$ . HiViT will process the features of the visible mask-units stage by stage, and finally get  $M' \times 512$  vectors as the output of encoder and feed them to the decoder for pixel restoration.

During the pre-training process, all three stages only need to process visible mask-units – this is why our model preserves the hierarchical structure and has efficient self-supervised pre-training efficiency.

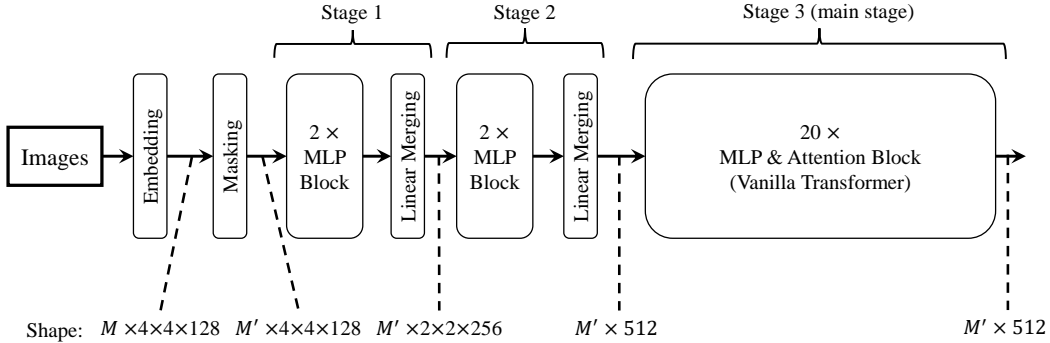


Figure 3: Pipeline of HiViT in MIM pre-training.

## References

- [1] Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.
- [2] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems*, 33:9912–9924, 2020.
- [3] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. *arXiv preprint arXiv:2104.14294*, 2021.
- [4] Chun-Fu Chen, Quanfu Fan, and Rameswar Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification. In *ICCV*, 2021.
- [5] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu

- Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. Mmdetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019.
- [6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.
- [7] 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*, 2022.
- [8] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised visual transformers. *arXiv preprint arXiv:2104.02057*, 2021.
- [9] Xiangxiang Chu, Zhi Tian, Bo Zhang, Xinlong Wang, Xiaolin Wei, Huaxia Xia, and Chunhua Shen. Conditional positional encodings for vision transformers. *Arxiv preprint 2102.10882*, 2021.
- [10] Zihang Dai, Hanxiao Liu, Quoc Le, and Mingxing Tan. Coatnet: Marrying convolution and attention for all data sizes. *Advances in Neural Information Processing Systems*, 34, 2021.
- [11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. Imagenet: A large-scale hierarchical image database. In *IEEE CVPR*, pages 248–255. IEEE CS, 2009.
- [12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*, 2019.
- [13] Xiaoyi Dong, Jianmin Bao, Dongdong Chen, Weiming Zhang, Nenghai Yu, Lu Yuan, Dong Chen, and Baining Guo. Cswin transformer: A general vision transformer backbone with cross-shaped windows. *arXiv preprint arXiv:2107.00652*, 2021.
- [14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.
- [15] Yuxin Fang, Li Dong, Hangbo Bao, Xinggang Wang, and Furu Wei. Corrupted image modeling for self-supervised visual pre-training. *CoRR*, abs/2202.03382, 2022.
- [16] Peng Gao, Jiasen Lu, Hongsheng Li, Roozbeh Mottaghi, and Aniruddha Kembhavi. Container: Context aggregation networks. In *NeurIPS*, pages 19160–19171, 2021.
- [17] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *ICLR*, 2018.
- [18] 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.
- [19] Kai Han, An Xiao, Enhua Wu, Jianyuan Guo, Chunjing Xu, and Yunhe Wang. Transformer in transformer. *NeurIPS*, 34, 2021.
- [20] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. *arXiv preprint arXiv:2111.06377*, 2021.
- [21] 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*, pages 9729–9738, 2020.
- [22] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask r-cnn. In *IEEE ICCV*, 2017.
- [23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE CVPR*, pages 770–778, 2016.
- [24] Zhouyuan Huo, Bin Gu, and Heng Huang. Large batch optimization for deep learning using new complete layer-wise adaptive rate scaling. In *AAAI*, pages 7883–7890, 2021.
- [25] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1106–1114, 2012.
- [26] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.
- [27] Chunyuan Li, Jianwei Yang, Pengchuan Zhang, Mei Gao, Bin Xiao, Xiyang Dai, Lu Yuan, and Jianfeng Gao. Efficient self-supervised vision transformers for representation learning. *arXiv preprint arXiv:2106.09785*, 2021.
- [28] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755. Springer, 2014.

- [29] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *ECCV*, pages 740–755, 2014.
- [30] 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. *arXiv preprint arXiv:2103.14030*, 2021.
- [31] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021.
- [32] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.
- [33] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *ECCV*, pages 69–84. Springer, 2016.
- [34] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *CVPR*, pages 2536–2544, 2016.
- [35] Zhiliang Peng, Wei Huang, Shanzhi Gu, Lingxi Xie, Yaowei Wang, Jianbin Jiao, and Qixiang Ye. Conformer: Local features coupling global representations for visual recognition. *arXiv preprint arXiv:2105.03889*, 2021.
- [36] 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*, pages 8821–8831. PMLR, 2021.
- [37] Jeong-geun Song. Ufo-vit: High performance linear vision transformer without softmax. *arXiv preprint arXiv:2109.14382*, 2021.
- [38] Aravind Srinivas, Tsung-Yi Lin, Niki Parmar, Jonathon Shlens, Pieter Abbeel, and Ashish Vaswani. Bottleneck transformers for visual recognition. In *CVPR*, pages 16519–16529, 2021.
- [39] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *ICML*, pages 6105–6114, 2019.
- [40] Yunjie Tian, Lingxi Xie, Jiemin Fang, Mengnan Shi, Junran Peng, Xiaopeng Zhang, Jianbin Jiao, Qi Tian, and Qixiang Ye. Beyond masking: Demystifying token-based pre-training for vision transformers. *CoRR*, abs/2203.14313, 2022.
- [41] Yunjie Tian, Lingxi Xie, Xiaopeng Zhang, Jiemin Fang, Haohang Xu, Wei Huang, Jianbin Jiao, Qi Tian, and Qixiang Ye. Semantic-aware generation for self-supervised visual representation learning. *CoRR*, abs/2111.13163, 2021.
- [42] Ilya O Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, et al. Mlp-mixer: An all-mlp architecture for vision. *NeurIPS*, 34, 2021.
- [43] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *ICML*, volume 139, pages 10347–10357, 2021.
- [44] Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. Going deeper with image transformers. *arXiv preprint arXiv:2103.17239*, 2021.
- [45] Ashish Vaswani, Prajit Ramachandran, Aravind Srinivas, Niki Parmar, Blake Hechtman, and Jonathon Shlens. Scaling local self-attention for parameter efficient visual backbones. In *CVPR*, pages 12894–12904, 2021.
- [46] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- [47] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pvt2: Improved baselines with pyramid vision transformer. *arXiv preprint arXiv:2106.13797*, 2021.
- [48] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *ICCV*, 2021.
- [49] Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. *arXiv preprint arXiv:2112.09133*, 2021.
- [50] Chen Wei, Lingxi Xie, Xutong Ren, Yingda Xia, Chi Su, Jiaying Liu, Qi Tian, and Alan L Yuille. Iterative reorganization with weak spatial constraints: Solving arbitrary jigsaw puzzles for unsupervised representation learning. In *CVPR*, pages 1910–1919, 2019.

- [51] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. Cvt: Introducing convolutions to vision transformers. *arXiv preprint arXiv:2103.15808*, 2021.
- [52] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *ECCV*, pages 418–434, 2018.
- [53] Zhenda Xie, Yutong Lin, Zheng Zhang, Yue Cao, Stephen Lin, and Han Hu. Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16684–16693, 2021.
- [54] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. *arXiv preprint arXiv:2111.09886*, 2021.
- [55] Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Xiyang Dai, Bin Xiao, Lu Yuan, and Jianfeng Gao. Focal self-attention for local-global interactions in vision transformers. *arXiv preprint arXiv:2107.00641*, 2021.
- [56] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. *arXiv preprint arXiv:2101.11986*, 2021.
- [57] Pengchuan Zhang, Xiyang Dai, Jianwei Yang, Bin Xiao, Lu Yuan, Lei Zhang, and Jianfeng Gao. Multi-scale vision longformer: A new vision transformer for high-resolution image encoding. In *ICCV*, 2021.
- [58] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *ECCV*, pages 649–666. Springer, 2016.
- [59] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *CVPR*, pages 633–641, 2017.
- [60] Daquan Zhou, Bingyi Kang, Xiaojie Jin, Linjie Yang, Xiaochen Lian, Zihang Jiang, Qibin Hou, and Jiashi Feng. Deepvit: Towards deeper vision transformer. *arXiv preprint arXiv:2103.11886*, 2021.