SparseMAE: Sparse Training Meets Masked Autoencoders

Aojun Zhou[1*] Yang Li[2] Zipeng Qin[1] Jianbo Liu[1] Junting Pan[1]
Renrui Zhang[13] Rui Zhao[2] Peng Gao[3] Hongsheng Li[1]
[1]The Chinese University of Hong Kong
[2]SenseTime Research [3]Shanghai AI Lab

Abstract

Masked Autoencoders (MAE) and its variants have proven to be effective for pretraining large-scale Vision Transformers (ViTs). However, small-scale models do not benefit from the pretraining mechanisms due to limited capacity. Sparse training is a method of transferring representations from large models to small ones by pruning unimportant parameters. However, naively combining MAE finetuning with sparse training make the network task-specific, resulting in the loss of task-agnostic knowledge, which is crucial for model generalization. In this paper, we aim to reduce model complexity from large vision transformers pretrained by MAE with assistant of sparse training. We summarize various sparse training methods to prune large vision transformers during MAE pretraining and finetuning stages, and discuss their shortcomings. To improve learning both task-agnostic and task-specific knowledge, we propose SparseMAE, a novel two-stage sparse training method that includes sparse pretraining and sparse finetuning. In sparse pretraining, we dynamically prune a small-scale sub-network from a ViT-Base. During finetuning, the sparse sub-network adaptively changes its topology connections under the task-agnostic knowledge of the full model. Extensive experimental results demonstrate the effectiveness of our method and its superiority on small-scale vision transformers. Code will be available at `https://github.com/aojunzz/SparseMAE`.

1. Introduction

Recently, several pretrained Masked Image Modeling (MIM) methods, such as MAE [15] and data2vec [1], have achieved great success in various computer vision tasks. They usually involve a two-stage training scheme where the model learns *task-agnostic knowledge* through a pretraining pretext task, and is subsequently finetuned on downstream tasks to acquire *task-specific knowledge*. Among these models, Masked Autoencoders (MAE) demonstrate superior learning capability on large-scale vision transformers (Fig. 1 (a)), thanks to large models' strong capacity to learn powerful general representations in the pretraining phase and versatile transferability to specific vision tasks. Yet, the large deep models are burdensome and difficult to be deployed to computationally restricted real-world scenarios. On the other hand, smaller and more efficient models, such as ViT-Tiny and ViT-Small, fail to perform well after pretrained by MAE or data2vec frameworks. For example, ViT-Tiny's finetuning results are even inferior to fully-supervised training (see Fig. 1 (b)). These empirical findings suggest that, under MIM frameworks, when the model is scaled down, it becomes more difficult to learn well via masked pretraining due to its limited capacity during pretraining and consequently hinders its performance when transferred to downstream tasks. In the realm of MIM pretraining, model capacity becomes a key ingredient to learning task-agnostic knowledge via the pretext tasks.

In our studies, we try to tackle the issues of MIM pretraining for small-scale vision transformers. Particularly, we apply pruning techniques [28, 20, 4, 42, 23] in the MIM frameworks in order to obtain performant small-scale models from larger ones. As a first step in our exploration, we use existing sparse training methods [42, 23] for MAE pretraining and finetuning on unstructured and hardware-friendly N:M sparsity to obtain a model with a similar scale to ViT-Tiny, which is pruned from ViT-Base (see Tab. 5 row 3). It achieves notably improvements compared to directly training ViT-Tiny using MAE on downstream ImageNet finetuning accuracy (77.6% vs. 71.6%) and ADE20K [43] semantic segmentation (39.8 mIoU vs. 37.6 mIoU). The empirical results indicate that on small-scale ViTs sparse training via pruning has promising applications under the MIM pretraining paradigm. However, there is still a large performance gap between the sparse model and the vanilla ViT-Base trained densely under MAE in their ImageNet finetuning accuracies (77.6% vs. 83.2%). The sparsity constraints adversely limit the model capacity and prevent it from learning good task-agnostic knowledge, similar to the predicament faced

*The first two authors equally contribute to this paper.

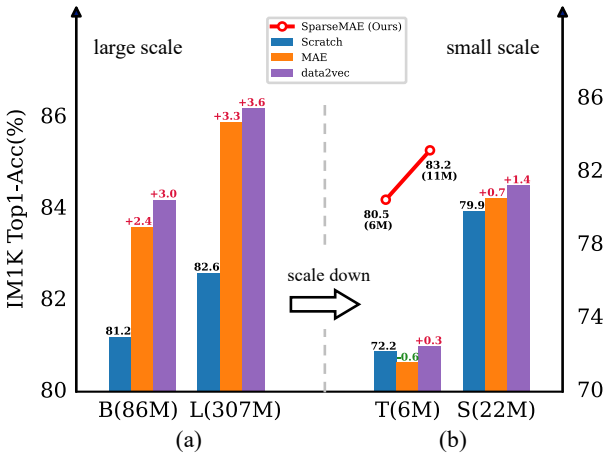


Figure 1: **(a)** Top-1 accuracy of large-scale models on ImageNet-1K, showing that large-scale vision transformers and self-supervised training can bring significant improvement. **(b)** On tiny-scale and small-scale, our SparseMAE outperforms other self-supervised methods, such as data2vec [1], MAE [15], and training from scratch.

by ViT-Tiny when it is directly pretrained under MAE.

To obtain more powerful small-scale models, we propose SparseMAE, a simple and unified sparse training method for Masked Autoencoders. Specifically, SparseMAE trains an adaptive sub-network in conjunction with its corresponding large dense model during pretraining. Both the large model and the sparse sub-network predict the same reconstruction objective and are optimized end-to-end. The large full model provides a strong capacity to learn task-agnostic knowledge, while the sparse sub-network can better preserve the knowledge by rewiring the connections in the full model. To better learn task-specific knowledge in downstream tasks, the sparse sub-network inherits the task-agnostic knowledge from the pretraining stage and also dynamically adjusts its connections during the finetuning phase.

We conduct extensive experiments on a variety of tasks, including image classification, object detection, segmentation, and robustness evaluation, to thoroughly evaluate the effectiveness of our proposed approach. Notably, our method achieves a top-1 accuracy of 80.5% with only 6M parameters on ImageNet-1K [8] classification and obtains a 45.2 mIoU on ADE20K segmentation with UperNet [37], significantly outperforming all previous works that have developed small-scale vision transformers using MAE (e.g., MAE-lite, TinyMIM and G2SD distillation).

In all, our contributions can be summarized in three folds:

- We thoroughly study the integration of sparse training as a pruning technique into the MAE framework. Specifically, we design three different strategies to train sparse networks under MAE and analyze their limitations.

- Based on the findings from above, we propose a novel sparse training framework, SparseMAE for Masked Autoencoders to improve the acquisition of both the task-agnostic and task-specific knowledge during the pretraining and finetuning stages.

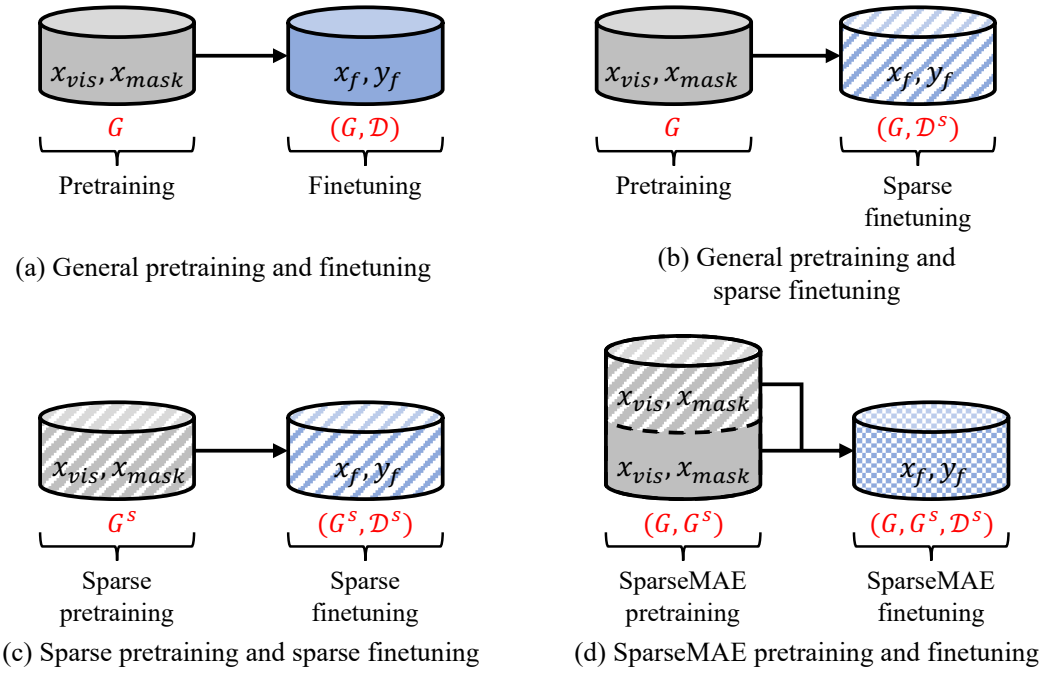
- We present extensive experiments to validate the effectiveness of our proposed method and achieve state-of-the-art performance with small-scale vision transformers on ImageNet classification and various downstream tasks.

2. Related works

2.1. Masked Image Modeling

Unsupervised pretraining on large-scale images with Masked Image Modeling (MIM) has shown superior performance on various computer vision tasks. BEiT [2] explores Masked Image Modeling on vision transformers by reconstructing the vision dictionary [30]. MAE [15] further proposes an asymmetric encoder and decoder for scaling up MIM to huge models. Besides, it demonstrates a simple pixel reconstruction loss can learn good visual representations. Due to the simplicity and computational efficiency, MAE is raising to a popular generative pretraining paradigm [40, 39]. As MAE reconstructs low-level signals with an isotropic vision transformer architecture, researchers improve MAE by exploring high-level signals architectures [1, 35, 12], which are more effective than reconstructing low-level signals.

However, those methods demonstrate the large capacity model is essential for good representation learning through the MIM pretext tasks. MAE-lite [34] explores Masked Autoencoders to improve the performance of tiny-scale vision transformer models, and introduces knowledge distillation into the pretraining phase. Different from MAE-lite, TinyMIM [31] proposes sequential relation knowledge distillation to improve the performance of small-scale vision transformer models, which utilize the full set of patch tokens and distill the knowledge from a pretrained ViT-Large model. TinyMIM [31] needs the large pretrained vision transformer and the use of sequential models is time-consuming for practical applications. These methods attempt to utilize the task-agnostic knowledge from large models through knowledge distillation. Different from their approaches, we explore sparse training as a pruning technique in order to scale down ViTs from large-scale to small-scale under the Masked Autoencoders framework.



(a) General pretraining and finetuning

(b) General pretraining and sparse finetuning

(c) Sparse pretraining and sparse finetuning

(d) SparseMAE pretraining and finetuning

Figure 2: An illustration of four types pretraining and finetuning frameworks. **(a)** Dense pretraining and finetuning the large full model as done in MAE [15]. G and D are general task-agnostic vision knowledge and task-specific downstream knowledge learnt from pretraining and finetuning respectively. **(b)** Learning G during dense pretraining with the large full model and acquiring downstream task-specific knowledge D^s for the sparse sub-network during sparse finetuning. **(c)** Learning task-agnostic knowledge G^s and task-agnostic knowledge D^s with sparse sub-network during sparse pretraining and finetuning. For finetuning, the sparse mask can be kept as fixed or dynamic. **(d)** Our SparseMAE training framework learns both G and G^s during pretraining and subsequently learns better D^s during finetuning.

2.2. Pruning and Sparse Training

Pruning and sparse training are critical methods for reducing the inference costs of deep neural networks [18]. Post-pruning methods [13, 26] typically involve training a dense full network first, followed by one or several rounds of train-prune-retrain rounds. Post-pruning methods are time-consuming due to the iterative manner [11]. Different from post-training methods, sparse training methods [29, 36] aim to adaptively obtain a sparse sub-network while optimizing both model weights and sparse connections from random initialization.

Sparse training is a class of methods stemming from the Sparse Evolutionary Training (SET) algorithm [29]. It starts from a randomly sparsified network and subsequently prunes and grows connections during training. These methods can be applied end-to-end within the network training stage and have achieved promising results on various vision benchmarks with more efficient models by adaptively changing the model topology connections during training.

Sparse training methods have shown promising results in various computer vision benchmarks, achieving state-of-the-art performance on different types of sparsity granularities such as unstructured sparsity [4, 23, 27], N:M fine-grained structured sparsity [42, 32, 10, 41], and block sparsity [7]. However, to our knowledge, the studies of pruning techniques for Vision Transformers [4, 20] only focus on the fully-supervised settings. There still lacks a unified method to prune large-scale Vision Transformers under the unsupervised Masked Autoencoders framework.

3. Methods

In this section, we first revisit the sparse training technique and Masked Autoencoders (MAE) in Sec. 3.1. Then we emphasize the challenges of integrating existing sparse training methods into MAE in Sec. 3.2. Lastly, we introduce our proposed method to effectively obtain a powerful small-scale sub-network from a large-scale model using sparse training in Sec. 3.3

3.1. Preliminaries

Revisiting Sparse Training. The pioneering work of Lottery Ticket Hypothesis (LTH) [11] demonstrates the dense

networks contain sparse matching subnetworks capable of training in isolation to full accuracy. Its findings inspire searching for strong sub-networks within a larger full network. Sparse training starts from randomly initializing the network weights $\mathbf{W}$ and associating them with a sparse mask $\mathbf{M}$, which is a binary tensor with 0 indexing the sparsified entries and 1 indicating the retained ones. $\mathbf{M}$ is optimized end-to-end and is usually obtained by magnitude pruning [14]. We denote the sparse sub-network as $\widetilde{\mathbf{W}} = \mathbf{W} \odot \mathbf{M}$. Sparse-to-sparse training regime optimizes the sparse weights $\widetilde{\mathbf{W}}$ and the corresponding binary mask $\mathbf{M}$, which can be updated at each iteration or periodically. Variants of sparse training, such as unstructured sparsity and fine-grained structured sparsity [23, 25, 43, 9] achieve state-of-the-art performance on various convolutional networks with improved efficiency.

Masked Autoencoders. Masked Autoencoders are a family of self-supervised methods for pretraining Vision Transformers (ViT). They involve reconstructing masked RGB patches from visible tokens x_{vis} with an encoder f and a decoder g. Mathematically, we have

$$\mathcal{L}_{\text{MAE}}(\mathbf{W}) = \mathbb{E}(\|g_\theta(f(x_{\text{vis}}, \mathbf{W})) - x_{\text{mask}}\|^2). \quad (1)$$

Here, an encoder f maps inputs x_{vis} to latent features $z_{\text{vis}} = f(x_{\text{vis}})$, and a decoder g_θ parameterized by θ predicts the masked token x_{mask} in the RGB pixel space from the latent feature z_{vis}. Encoder weights $\mathbf{W}$ and decoder weights θ are jointly trained. After pretraining, the decoder is discarded and the encoder is finetuned on task-specific downstream datasets. During finetuning, the model acquires downstream specific knowledge D, which is built on general task-agnostic knowledge G learnt in pretraining, as shown in Fig. 2 (a).

3.2. Challenges of Sparse Training for MAE

Despite the effectiveness of MAE for large-scale vision transformers, it encounters great difficulties when applied to smaller models (see Fig. 1). The inadequate model capacity prevents small models from learning meaning task-agnostic knowledge and consequently transfers to inferior performance in downstream tasks. To enable the potential of large models' capability for smaller ones to learn powerful representations under MAE framework, we can modify Eq. 1 to combine it with sparse training as follows:

$$\mathcal{L}_{\text{Sparse}}(\mathbf{W}, \mathbf{M}) = \mathbb{E}(\|g_\theta(f(x_{\text{vis}}, \mathbf{W} \odot \mathbf{M})) - x_{\text{mask}}\|^2), \quad (2)$$

where f represents the large full encoder with parameters $\mathbf{W}$. The encoder binary mask $\mathbf{M}$ produces a sparse sub-network by magnitude pruning during pretraining, while the decoder is kept dense as it is not used for downstream finetuning. Following the notation in Sec. 3.1, we use G^s to represent the general knowledge of the sparse sub-network learnt in pretraining, and we denote D^s the specific knowledge of the finetuned sparse sub-network for downstream tasks. Considering both the pretraining and finetuning phases, we outline three intuitive strategies to integrate sparse training into the MAE framework:

- 1 Dense pretraining then sparse finetuning as shown in Fig. 2 (b). Although this strategy allows the sparse finetuning phase to optimize from the large full model's task-agnostic knowledge G, the sub-network does not benefit from the pretraining task and lacks its knowledge G^s. This may hinder the model to acquire better downstream knowledge D^s
- 2&3 Sparse pretraining using Eq. (2) then sparse finetuning (see Fig. 2 (c)). During finetuning, the sparsity mask can be kept as fixed or dynamic (dynamic mask is the third strategy). This strategy relies on the architecture and task-agnostic knowledge G^s obtained by the sub-network during the pretraining phase, while it ignores the potential benefits brought by incorporating the knowledge G of the large full model.

In summary, it is challenging to tailor a sparse training mechanism for MAE framework to sufficiently learn task-agnostic and task-specific knowledge for a small model. In the following section, we introduce our approach to solve this dilemma by fully utilizing the task-agnostic knowledge G of the large full model and adaptively optimizing sub-network's topology connections for various downstream tasks for better task-specific knowledge.

3.3. Sparse Masked Autoencoders

To tackle the limitations of the sparse training strategies introduced above, we propose a novel method for pruning ViT models under the MAE framework (see Fig. 2 (d)).

Firstly, to ensure a good general task-agnostic knowledge G to be learnt via pretraining, we utilize the dense full model to reconstruct the RGB pixels following Eq. 1, which learns more discriminative representations thanks to its larger capacity. Simultaneously, a sparse sub-network within the full dense network is trained using the same visible tokens as the inputs and regresses the same targets as the large full model. The sparseMAE loss function is therefore extended from Eq. 1 and Eq. 2 as follows:

$$\mathcal{L} = \underbrace{\mathcal{L}_{\text{Sparse}}(\widetilde{\mathbf{W}})}_{\text{sub-network}} + \underbrace{\alpha \mathcal{L}_{\text{MAE}}(\mathbf{W})}_{\text{full network}}, \quad (3)$$

where $\mathbf{W}$ represents the dense parameters of the large full model, $\widetilde{\mathbf{W}} = \mathbf{W} \odot \mathbf{M}$ is the masked sparse parameters of the sub-network, and α is the scaling hyperparameter balancing the reconstruction losses for the dense full model and the

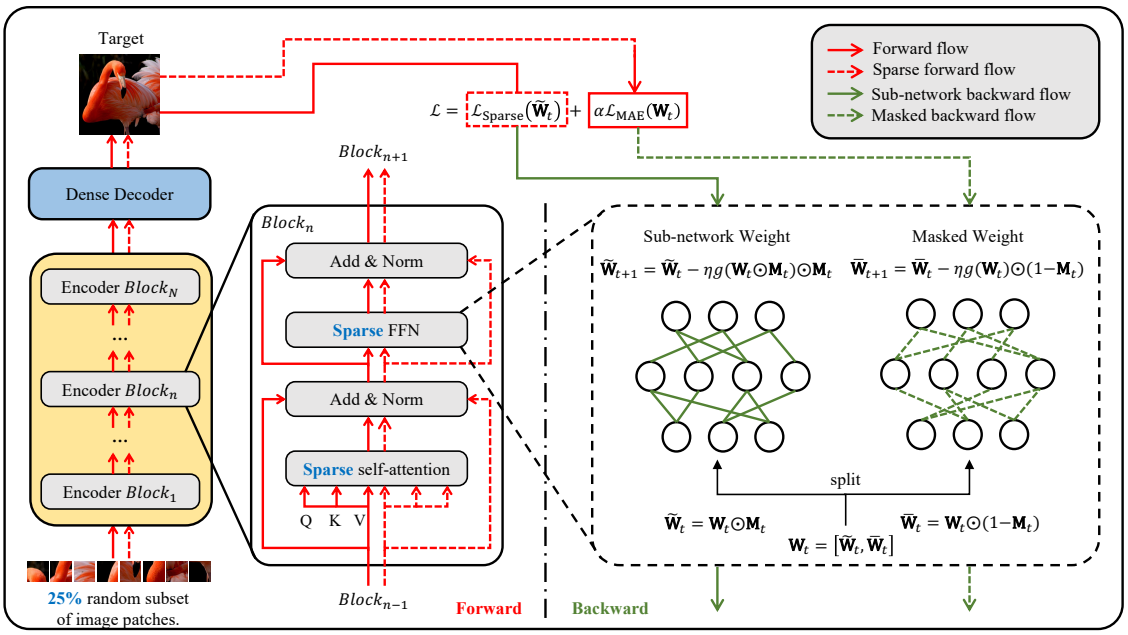


Figure 3: Illustration of Sparse Masked Autoencoders. In the forward pass, both the large full model (red solid lines) and the sparse sub-network (red dashed lines) reconstruct the masked patches from the same visible tokens. In the backward pass, the masked weights (green solid lines) and the sparse weights (green dashed lines) are updated separately using the gradients from the corresponding losses.

sparse sub-network. In our experiments, we set $\alpha = 1.0$ by default.

As shown in Eq. 3, the gradients for the sparse weights $\widetilde{\mathbf{W}}$ shall be updated following both the reconstruction losses of the sub-network and the full network, which may lead to gradient imbalance between the masked and unmasked weights [38]. Therefore, in each training iteration we split the weights $\mathbf{W}$ of the full model into sparse sub-network weights $\widetilde{\mathbf{W}}$ and unmasked weights $\overline{\mathbf{W}} = \mathbf{W} \odot (1 - \mathbf{M})$ and optimize them separately:

$$\widetilde{\mathbf{W}}_{t+1} = \widetilde{\mathbf{W}}_t - \eta(g(\mathbf{W}_t \odot \mathbf{M}_t) \odot \mathbf{M}_t), \tag{4}$$

$$\overline{\mathbf{W}}_{t+1} = \overline{\mathbf{W}}_t - \eta(g(\mathbf{W}_t) \odot (1 - \mathbf{M}_t)), \tag{5}$$

where $g(\mathbf{W}_t \odot \mathbf{M}_t) = \frac{\partial \mathcal{L}_{\text{Sparse}}(\mathbf{W}_t \odot \mathbf{M}_t)}{\partial \mathbf{W}_t}$ are gradients of the sparse sub-network, $g(\mathbf{W}_t) = \frac{\partial \mathcal{L}_{\text{MAE}}(\mathbf{W}_t)}{\partial \mathbf{W}_t}$ are the gradients of the full models, η is the learning rate and t indexes the training iteration. The binary mask $\mathbf{M}_t$ is obtained by on-the-fly magnitude pruning. Here, we only show gradient descent for simplicity and omit other parts (e.g., weight decay). In this way, the large full model can utilize all the model parameters in the forward pass to produce discriminative outputs, while in the backward pass, the sparse sub-network stays unaffected by the gradients from the large full model. During the pretraining phase, the task-agnostic knowledge G^s of the sparse sub-network can dynamically select the important sparse connections in the context of the task-agnostic knowledge G of the full dense model, which helps to identify the important sparse topology and is retained for downstream finetuning.

After pretraining, the full dense model with a sparse sub-network is transferred and finetuned on different downstream tasks (e.g., classification, detection, segmentation, etc.). For the finetuning stage, the deep representations from the sparse sub-network are used for the downstream tasks, while its topology connections are optimized end-to-end given the specific downstream objectives. We refrain from fine-tuning dense full models, as the fine-tuning process for both dense and sparse models is costly. Hence, the learning of the task-specific knowledge D^s for downstream tasks can benefit from inheriting a pretrained sparse sub-network with knowledge G^s, and adaptively selecting the more suitable sparse architecture built on the general knowledge G of the full model.

4. Experiments

In this section, we validate the effectiveness and robustness of our SparseMAE. First, we introduce the implementation details of our SparseMAE in Sec. 4.1. Then, we present

our main experimental results in Sec. 4.2, which include classification results on ImageNet-1K [8] and semantic segmentation results on ADE20K [43]. To demonstrate the superiority of our proposed method, we compare SparseMAE with other state-of-the-art ViT pruning methods in Sec. 4.4. In Sec. 4.3, we evaluate the transferring ability of the models trained with our SparseMAE to object detection, instance segmentation tasks on MS COCO [24] and robustness evaluation. Finally, we perform ablative experiments to evaluate the performance of SparseMAE under different settings in Sec. 4.5.

4.1. Implementation Details

Sparse Training Methods. We conduct experiments with two different granularities for sparse training, consisting of unstructured sparsity [14, 21] and hardware-friendly N:M sparsity [28, 42]. As we focus on studying how to learn and transfer task-agnostic and task-specific knowledge under the Masked Autoencoders paradigm with sparse training methods, we select two simple and effective sparse training methods, namely DFP [23] for unstructured sparsity and SR-STE [42] for N:M sparsity. We prune all the linear layers within the vision transformers and set the relevant hyperparameters for sparse training the same as [23, 42] by default.

Models. For all of our main experiments, unless otherwise stated, we select ViT-Base as our large full model and prune it for smaller ones. For hardware-friendly N:M sparsity, we use 2:32 and 4:32 sparse ratios to obtain models with similar scales to ViT-Tiny (5.8M) and ViT-Small (11.3M), *the main experiments use N:M sparsity for default, which is applied to all linear layers, including self.qkv, self.proj in the Attention block, and self.fc1 and self.fc2 in the MLP block.* For unstructured sparsity, we extend the above N:M ratios and prune ViT-Base with 93.5% and 87.0% sparsity. We name our ViT-Base with 2:32/93.5% sparsity as SparseMAE-Tiny (SparseMAE-T) and the one with 4:32/87.0% sparsity as SparseMAE-Small (SparseMAE-S).

Pretraining. Following MAE [15] pretraining scheme, models start from *random initialization*, we use ImageNet-1K (IN-1K) [8] for all pretraining experiments. We pretrain SparseMAE for 400 epochs using 224×224 images as inputs. The other settings also follow those used in [15]. For example, We use random resized cropping and horizontal flipping for data augmentation, a 75% mask ratio during pretraining, and a decoder with a single transformer layer and 512 hidden dimensions. Unlike other methods [34, 19] for small and tiny-scale vision transformers pretraining, we do not grid-search the decoder design, possible putting our method at a disadvantage.

Finetuning. We evaluate our SparseMAE pretrained models for various downstream tasks, including ImageNet-1K [8] image classification, MS COCO [24] object detection and instance segmentation tasks, ADE20K [43] semantic segmentation.

For image classification, we take a ViT-Base model pretrained using MAE for 400 epochs as the reference model, which is implemented using the officially released codebase [15] and achieves 83.2% top-1 accuracy. We finetune SparseMAE-Tiny/Small for 200 epochs using the same receipt described in [15] for a fair comparison with other works [31, 19] and set the batch size as 1024 and the base learning rate as 10^{-3}.

For semantic segmentation, we follow BEiT [2] to use UperNet[37] as our segmentation framework. We train the model with an input resolution of 512×512 for 160k iterations. A batch size of 16 is adopted. For object detection and instance segmentation tasks, we follow ViTDet [22] framework. We train our SparseMAE models with a batch size of 64 for 100 epochs, and the input image resolution is set to 1024×1024. Further details on the hyper-parameters can be found in Appendix.

4.2. Main Results

Classification and Semantic Segmentation. Tab. 1 shows the evaluation of our SparseMAE model against self-supervised methods for ViT-Tiny/Small on ImageNet-1K image classification and ADE20K semantic segmentation. For ImageNet-1K classification, SparseMAE achieves 80.5% top-1 accuracy. To our knowledge, this is currently the best result achieved by tiny-scale transformer models. Furthermore, SparseMAE outperforms TinyMIM [31]/G2SD [19]/MAE-lite [34] with +4.7/4.2/4.0 accuracy gains on tiny-scale transformers. For semantic segmentation, SparseMAE achieves an mIoU of 45.2, outperforming G2SD/MAE-lite/TinyMIM by +3.8/1.3/1.2 mIoUs. Similar improvements can be observed for the small-scale transformers. SparseMAE achieves a 83.2% ImageNet-1K classification accuracy and a 48.4% mIoU on ADE20K semantic segmentation, surpassing TinyMIM [31]/G2SD [19] by +0.2/1.2 and +0.0/2.2 on the two tasks respectively. The significant improvements over previous results show the superiority of our method in obtaining performant small models under MAE framework.

Speedups. Unstructured sparsity and N:M structured sparsity is efficient for commercial CPUs and FPGAs respectively. For our SparseMAE with unstructured sparsity, we tested the throughput (img/s) on an Intel RXeon(R) 6238R CPU using the deepsparse engine with a batch size of 64. For our SparseMAE with N:M sparsity, we tested the throughput on Xilinx XCVU13P with a batch size of 16. Our SparseMAE-S with N:M and unstructured sparsity achieved throughput gains of +71% and +9% over ViT-small with similar amount of parameters on FPGAs and CPUs. Therefore, our approach demonstrate high feasibility in actual scenarios.

Table 1: Finetuning results on ImageNet-1K and ADE20K. All our models are pre-trained only on ImageNet-1K. Extra pretrained teacher means the compared method cannot trained from scratch. ‡ the model adopts the MAE pretrained model before perform the sparse pretraining.

Method	Params (M)	Pretraining epochs	Encoder ratio	Extra Pretrained/ Teacher	Classification Top-1 Acc (%)	Segmentation mIoU (%)
DeiT [33]	5.8	300	-	Label	72.2	38.0
Tiny-scale						
MAE [15]	5.8	1600	25%	✗	71.6	37.6
MoCo [6]	5.8	300	100%	EMA	73.3	39.3
TinyMIM [31]	5.8	300	100%	TinyMIM-S	75.8	44.0
G2SD [19]	5.8	300	100%	ViT-B	76.3	41.4
MAE-lite [34]	5.8	400	25%	ViT-B	76.5	43.9
SparseMAE (ours)	5.8	400	25%	✗	**80.5**	**45.2**
DeiT [33]	22	300	-	Label	79.9	
Small-scale						
MAE [15]	22	1600	25%	✗	80.6	42.8
MoCo [6]	22	300	100%	EMA	81.4	43.9
DINO [3]	22	300	100%	EMA	81.5	45.3
CAE [5]	22	300	100%	DALL-E	82.0	-
TinyMIM [31]	22	300	100%	TinyMIM-ViT-B	83.0	48.4
G2SD [19]	22	300	25%	ViT-B	82.0	46.2
SparseMAE (ours)	11.3	300	25%	ViT-B‡	**83.2**	**48.4**
SparseMAE (ours)	11.3	400	25%	✗	82.1	46.7

4.3. Evaluation on Other Downstream Tasks

Object Detection and Instance Segmentation. We finetune the SparseMAE pretrained models on MS COCO with ViTDet [22]. As shown in Tab. 2, we report AP^{bbox} for object detection and AP^{mask} for instance segmentation. We compare SparseMAE-Tiny with MAE-Tiny and its improved variants using distillation methods, such as MAE-lite [34] and G2SD [19]. The results in Tab. 2 show that our SparseMAE obtains more than 3.1 AP^{bbox} and 2.4 AP^{mask} gains compared to MAE-lite and G2SD, validating the transferability of the learnt representations in our models.

Robustness Evaluation. In Tab. 3, we evaluate the robustness of our models on different variants of ImageNet validation sets. We use the same models finetuned on original ImageNet (Tab. 1) and only run inference on the different validation sets, such as Imagenet-R [17], ImageNet-A [17] and ImageNet-C [16]), without any specialized fine-tuning. Tab. 3 shows that our method has better generalization capability than distillation-based methods, such as G2SD [19] and TinyMIM [31]).

4.4. Comparison with State-of-the-art ViT Pruning

To the best of our knowledge, our proposed SparseMAE is the first to study network pruning technique under the Masked Autoencoders paradigm. To validate the effectiveness of SparseMAE, we compare it against recent state-of-the-art fully-supervised pruning methods for ViTs. We choose SViTE [4] and oViT [20] as our compared models, which exploit end-to-end sparse training for ViTs and achieve state-of-the-art results in the fully-supervised setting. SViT [4] adopts a dynamic sparse training method to prune the dense model using random initialization. However, this approach requires training the sparse model for

Table 2: Object detection and instance segmentation results on the MS COCO dataset.

Method	Backbone Size (M)	Det AP^{bbox}	Seg AP^{mask}
Tiny-scale models			
DeiT [33]	5.8	40.7	36.5
MAE [15]	5.8	38.9	35.1
MoCo v3 [6]	5.8	40.0	36.0
DINO [3]	5.8	40.2	36.1
MAE-lite [34]	5.8	42.7	38.2
G2SD [19]	5.8	44.0	39.6
SparseMAE (ours)	5.8	**47.1**	**42.0**

Table 3: Robustness evaluation on ImageNet variants (top-1 accuracy, except for ImageNet-C which evaluates mean corruption error). We test the small-scale and tiny-scale models from Tab. 1 on different ImageNet validation sets, without any fine-tuning.

Method	Model Size	IN-A↑	IN-R↑	IN-C ↓
DeiT [33]		8.0	32.7	54.0
MAE [15]		7.0	36.5	55.2
TinyMIM [31]	Tiny-scale	11.0	39.8	50.1
G2SD [19]		12.9	39.0	-
SparseMAE (ours)		**18.2**	**45.9**	**47.2**
DeiT [33]		18.3	42.3	41.4
MAE [15]		20.1	45.6	40.6
TinyMIM [31]	Small-scale	27.5	48.8	35.8
G2SD [19]		29.4	46.8	-
SparseMAE (ours)		**29.3**	**49.2**	**35.2**

Table 4: Comparison with state-of-the-art supervised Vision Transformer pruning methods on ImageNet-1K.

Model	Method	Params (M)	Top-1 Acc (%)
ViT-Base	SViTE [4]	43.3	81.5
	oViT [20]	43.3	81.6
	SViTE [4]	34.6	81.3
	oViT [20]	34.6	81.5
	oViT [20]	21.6	81.1
	oViT [20]	17.3	80.8
	oViT [20]	8.7	79.7
ViT-Base	SparseMAE-T (ours)	**5.8**	**80.5**
	SparseMAE-S (ours)	**11.3**	**83.2**

600 epochs. oViT [20] (post-training pruning) prunes the model initialized from a trained dense model (300 epochs), which requires 300 epochs of finetuning. For our SparseMAE, we only need 400 epochs for pretraining and 200 epochs for fine-tuning. The pretraining of SparseMAE is 3× faster than full-token supervised training. Consequently, our SparseMAE approach reduces the training time from 968 GPU hours to 581 GPU hours compared to SOTA SViT and oViT.

Tab. 4 demonstrates that SparseMAE can outperform state-of-the-art fully-supervised pruning counterparts significantly for small and tiny-scale transformers. Our SparseMAE-T has a 0.8% performance gain and a 35.0% computation complexity reduction compared to oViT [20] for tiny-scale transformers. With our SparseMAE-S, we achieve a 83.2% ImageNet classification accuracy, outperforming supervised pruning techniques by large margins. These results suggest that our method successfully extends the superiority of MAE for representation learning with large-scale vision transformers to small and tiny-scale models.

Table 5: Different sparse pretraining or finetuning strategies with tiny-scale models on ImageNet-1K and ADE20K. "S" means sparse pretraining, "D" means dense pretraining, and "D+S" means sparse pretraining together with a dense full model pretraining. ‡: the model is finetuned with the fixed sparse masks from pretraining. Corr. Strategy.: corresponding sparse training strategies explained in Sec. 4.5.1.

Model	Pretrain Method	Top-1 Acc (%)	Seg mIoU (%)	Corr. Strtg.
ViT-Base	D	83.2	48.1	-
Unstructured sparsity (93.5% sparse ratio)				
SparseMAE-T	D	77.8	40.2	#1
SparseMAE-T	S‡	76.7	37.7	#2
SparseMAE-T	S	77.6	39.8	#3
SparseMAE-T	D+S	**80.4**	**45.4**	-
2:32 sparsity				
SparseMAE-T	D	78.4	40.7	#1
SparseMAE-T	S‡	76.3	37.2	#2
SparseMAE-T	S	78.7	41.3	#3
SparseMAE-T	D+S	**80.5**	**45.2**	-

4.5. Ablation Studies and Discussion

4.5.1 Comparison of Sparse Training Strategies.

We compare three sparse training strategies for MAE framework introduced in Sec. 3.2 under two sparse granularities. Namely, we (1) pretrain a large full model using MAE and finetune it with sparse training, (2) combine sparse training with MAE pretraining and subsequently finetune the model on downstream tasks with the fixed sparse mask from pretraining, or (3) modify (2) to use dynamic sparse mask for finetuning. The results are shown in Tab. 5. Under unstructured sparsity, comparing (1) and (2), we see that the classification performance drops from 77.8% to 76.7%, showing that the knowledge G^s learnt by the sparse sub-network during pretraining cannot generalize well to the downstream tasks compared to the task-agnostic knowledge G learnt by the large full model. (3) further allows the adaptation of sparse sub-network architecture during finetuning and partially alleviates the problem by better acquiring downstream specific knowledge D^s. However, its result still lags significantly behind our proposed method (77.6% vs. 80.4%), where the difference is only the absence of a concurrently pretrained large full model. On the other hand, when comparing (1) with our method, the addition of a pretrained sparse sub-network can boost the performance by 2.6%. Similar observations can also be made under N:M sparsity settings and semantic segmentation results. These empirical results suggest that both the task-agnostic knowledge G of the large full model and G^s of the sparse sub-network are necessary for small and tiny-scale vision transformers to effectively benefit from MAE pretraining.

Table 6: Comparison between different full models for SparseMAE on ImageNet-1K and ADE20K.

Model	Pattern	Params (M)	Top-1 Acc (%)	Seg mIoU (%)
ViT-Small	8:32	5.8	77.4	41.5
ViT-Base	2:32	5.8	**80.5**	**45.2**

4.5.2 Ablation of Full Model Capacity

We conduct an ablation study to investigate the performance with different capacities of the full model. We use our SparseMAE and set ViT-Small and ViT-Base as the full models, where ViT-Small only has 25% parameters of ViT-Base. As shown in Tab. 6, the results show that our SparseMAE in ViT-Base achieves 3.1% top-1 accuracy advantage over ViT-Small, indicating that larger capacity of the full model with SparseMAE can significantly improve its performance by incorporating better pretraining task-agnostic knowledge.

4.5.3 Impact of Different Pruning Sparsity Ratios

The main experiments of sparse ratio setting is compared to the ViT-tiny and ViT-small. We provide additional sparse ratios to verify the performance of SparseMAE on ImageNet-1K as in Tab. 7.

Table 7: Comparison between different sparse ratios for SparseMAE on ImageNet-1K.

Sparse Pattern	1:32	2:32	4:32	8:32
Top-1 Acc	76.8	80.5	83.2	83.7

5. Conclusion

In this paper, we introduce SparseMAE, an innovative method for incorporating sparse training into the Masked Autoencoders (MAE) framework to create efficient and powerful small and tiny-scale transformers. As the first study in this area, we first point out the limitations of sparse training under MAE and design a unified framework to transfer the task-agnostic and task-specific knowledge of large models to lightweight sparse sub-networks. SparseMAE trains the sparse sub-networks by persevering the task-agnostic knowledge of large full models and then adaptively finds the optimal specific sub-networks for downstream tasks. The proposed method outperforms the state-of-the-art small-scale vision transforms in both unsupervised pretraining and fully-supervised settings. We hope this work could shed light on sparse training with the Masked Autoencoders framework in the vision community.

Acknowledgement

This project is funded in part by National Key R&D Program of China Project 2022ZD0161100, by the Centre for Perceptual and Interactive Intelligence (CPII) Ltd under the Innovation and Technology Commission (ITC)'s InnoHK, by General Research Fund of Hong Kong RGC Project 14204021. Hongsheng Li is a PI of CPII under the InnoHK.

References

[1] Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. Data2vec: A general framework for self-supervised learning in speech, vision and language. In *International Conference on Machine Learning*, pages 1298–1312. PMLR, 2022.

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

[3] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.

[4] Tianlong Chen, Yu Cheng, Zhe Gan, Lu Yuan, Lei Zhang, and Zhangyang Wang. Chasing sparsity in vision transformers: An end-to-end exploration. *Advances in Neural Information Processing Systems*, 34:19974–19988, 2021.

[5] 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.

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

[7] Tri Dao, Daniel Y Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. *arXiv preprint arXiv:2205.14135*, 2022.

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[9] Utku Evci, Trevor Gale, Jacob Menick, Pablo Samuel Castro, and Erich Elsen. Rigging the lottery: Making all tickets winners. In *International Conference on Machine Learning*, pages 2943–2952. PMLR, 2020.

[10] Chao Fang, Aojun Zhou, and Zhongfeng Wang. An algorithm–hardware co-optimized framework for accelerating n: M sparse transformers. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 30(11):1573–1586, 2022.

[11] Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. *arXiv preprint arXiv:1803.03635*, 2018.

[12] Peng Gao, Renrui Zhang, Rongyao Fang, Ziyi Lin, Hongyang Li, Hongsheng Li, and Qiao Yu. Mimic before reconstruct: Enhancing masked autoencoders with feature mimicking. *IJCV 2023*, 2023.

[13] Song Han, Huizi Mao, and William J Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *arXiv preprint arXiv:1510.00149*, 2015.

[14] Song Han, Jeff Pool, John Tran, and William Dally. Learning both weights and connections for efficient neural network. *Advances in neural information processing systems*, 28, 2015.

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

[16] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *Proceedings of the International Conference on Learning Representations*, 2019.

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

[18] Torsten Hoefler, Dan Alistarh, Tal Ben-Nun, Nikoli Dryden, and Alexandra Peste. Sparsity in deep learning: Pruning and growth for efficient inference and training in neural networks. *The Journal of Machine Learning Research*, 22(1):10882–11005, 2021.

[19] Wei Huang, Zhiliang Peng, Li Dong, Furu Wei, Jianbin Jiao, and Qixiang Ye. Generic-to-specific distillation of masked autoencoders. *arXiv preprint arXiv:2302.14771*, 2023.

[20] Denis Kuznedelev, Eldar Kurtic, Elias Frantar, and Dan Alistarh. ovit: An accurate second-order pruning framework for vision transformers. *arXiv preprint arXiv:2210.09223*, 2022.

[21] Yann LeCun, John Denker, and Sara Solla. Optimal brain damage. *Advances in neural information processing systems*, 2, 1989.

[22] Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IX*, pages 280–296. Springer, 2022.

[23] Tao Lin, Sebastian U Stich, Luis Barba, Daniil Dmitriev, and Martin Jaggi. Dynamic model pruning with feedback. *arXiv preprint arXiv:2006.07253*, 2020.

[24] 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 *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[25] Shiwei Liu, Lu Yin, Decebal Constantin Mocanu, and Mykola Pechenizkiy. Do we actually need dense over-parameterization? in-time over-parameterization in sparse training. In *International Conference on Machine Learning*, pages 6989–7000. PMLR, 2021.

[26] Zhuang Liu, Mingjie Sun, Tinghui Zhou, Gao Huang, and Trevor Darrell. Rethinking the value of network pruning. *arXiv preprint arXiv:1810.05270*, 2018.

[27] Zhi-Gang Liu, Paul N Whatmough, and Matthew Mattina. Sparse systolic tensor array for efficient cnn hardware acceleration. *arXiv preprint arXiv:2009.02381*, 2020.

[28] Asit Mishra, Jorge Albericio Latorre, Jeff Pool, Darko Stosic, Dusan Stosic, Ganesh Venkatesh, Chong Yu, and Paulius Micikevicius. Accelerating sparse deep neural networks. *arXiv preprint arXiv:2104.08378*, 2021.

[29] Decebal Constantin Mocanu, Elena Mocanu, Peter Stone, Phuong H Nguyen, Madeleine Gibescu, and Antonio Liotta. Scalable training of artificial neural networks with adaptive sparse connectivity inspired by network science. *Nature communications*, 9(1):2383, 2018.

[30] 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.

[31] Sucheng Ren, Fangyun Wei, Zheng Zhang, and Han Hu. Tinymim: An empirical study of distilling mim pre-trained models. *arXiv preprint arXiv:2301.01296*, 2023.

[32] Wei Sun, Aojun Zhou, Sander Stuijk, Rob G. J. Wijnhoven, Andrew Nelson, Hongsheng Li, and Henk Corporaal. Dominosearch: Find layer-wise fine-grained n:m sparse schemes from dense neural networks. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021.

[33] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR, 2021.

[34] Shaoru Wang, Jin Gao, Zeming Li, Jian Sun, and Weiming Hu. A closer look at self-supervised lightweight vision transformers. *arXiv preprint arXiv:2205.14443*, 2022.

[35] Chen Wei, Haoqi Fan, Saining Xie, Chao Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. 2021.

[36] Mitchell Wortsman, Ali Farhadi, and Mohammad Rastegari. Discovering neural wirings. *Advances in Neural Information Processing Systems*, 32, 2019.

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

[38] Jiahui Yu, Linjie Yang, Ning Xu, Jianchao Yang, and Thomas Huang. Slimmable neural networks. *arXiv preprint arXiv:1812.08928*, 2018.

[39] Renrui Zhang, Ziyu Guo, Peng Gao, Rongyao Fang, Bin Zhao, Dong Wang, Yu Qiao, and Hongsheng Li. Point-m2ae: Multi-scale masked autoencoders for hierarchical point cloud pre-training. *NeurIPS 2022*, 2022.

[40] Renrui Zhang, Liuhui Wang, Yu Qiao, Peng Gao, and Hongsheng Li. Learning 3d representations from 2d pre-trained models via image-to-point masked autoencoders. *CVPR 2023*, 2023.

[41] Yuxin Zhang, Mingbao Lin, Zhihang Lin, Yiting Luo, Ke Li, Fei Chao, Yongjian Wu, and Rongrong Ji. Learning best combination for efficient n:m sparsity. *arXiv preprint arXiv:2206.06662*, 2022.

[42] Aojun Zhou, Yukun Ma, Junnan Zhu, Jianbo Liu, Zhijie Zhang, Kun Yuan, Wenxiu Sun, and Hongsheng Li. Learning n: M fine-grained structured sparse neural networks from scratch. *arXiv preprint arXiv:2102.04010*, 2021.

[43] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017.