TOWARDS SUSTAINABLE SELF-SUPERVISED LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Though increasingly training-expensive, most self-supervised learning (SSL) models have repeatedly been trained from scratch but not fully utilized since only a few SOTAs are adopted for downstream tasks. In this work, we explore a sustainable SSL framework with two major challenges: i) learning a stronger new SSL model based on the existing pretrained SSL model in a cost-friendly manner, ii) allowing the training of the new model to be compatible with various base models. We propose a Target-Enhanced Conditional (TEC) scheme, which introduces two components to existing mask-reconstruction based SSL. Firstly, we introduce patch-relation enhanced targets to encourage the new model to learn semantic-relation knowledge from the base model using incomplete inputs. This hardening and target-enhancing could help the new model surpass the base model, since they enforce additional patch relation modeling to handle incomplete input. Secondly, we introduce a conditional adapter that adaptively adjusts new model prediction to align with the target of each base model. Experimental results show that our TEC scheme can accelerate the learning speed and also improve SOTA SSL models, *e.g.,* MAE and iBOT, taking an explorative step towards sustainable SSL.

1 INTRODUCTION

Self-supervised learning (SSL) has achieved overwhelming success in unsupervised representation learning, with astonishing high performance in many downstream tasks like classification (Zhou et al., 2022a;b), object detection and segmentation (Bao et al., 2021; He et al., 2022). In SSL, a pretext task is first built, *e.g.,* instance discrimination task (He et al., 2020; Chen* et al., 2021) or masked image modeling (MIM) (Bao et al., 2021; He et al., 2022), and then pseudo labels are generated via the pretext task to train a network model without requiring manual labels. Though successful, SSL is developing towards a direction of causing increasingly large training costs, *e.g.,* MoCo (He et al., 2020) trained with 200 epochs while MAE (He et al., 2022) with 16,00 epochs to release its potential. However, in most cases researchers only have limited computational budgets and often cannot afford to train large SSL models. Moreover, the pretrained non-SOTA SSL models are rarely used in practice, since SOTA is updated frequently and a previous one quickly becomes useless, wasting huge training resources. Thus, a **sustainable SSL** framework is much demanded.

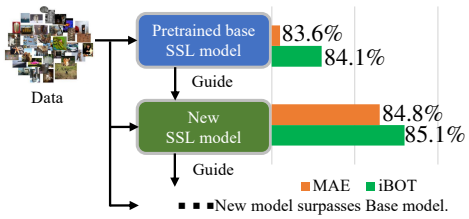

Figure 1: The concept of sustainable SSL.

Just like how human experience is enriched and passed from one generation to the next in human society, we try to make an SSL model absorb knowledge from the pretrained SSL model to achieve superior representation quality, so as to be reusable or "sustainable". In this way, huge training resources would be saved compared to training a new SSL model from scratch. This process is somewhat like Knowledge Distillation (KD) (Hinton et al., 2015; Gou et al., 2021). But instead of aiming at compressing knowledge from a powerful teacher model to a compact new model in traditional KD, usually with declined performance, the sustainable SSL is targeted at a new model that learns from the base model and becomes more powerful than the base model. An illustration of sustainable SSL is given in Fig. 1. For better understanding, we call the new SSL model to be trained as the new model and the pretrained SSL model as the base model. To surpass the base model, in sustainable SSL, the new model exploits not only the implicit base model knowledge but also the

absent knowledge in the base model. Such a learning process follows a fully self-supervised manner, and differs from the self-training schemes (Xie et al., 2020; Yalniz et al., 2019) that requires labels for supervised learning.

In this work, we take an explorative step towards sustainable SSL by learning from existing pretrained SSL models and surpassing them in an efficient manner. Intuitively, to achieve this challenging goal, the new model should learn not only knowledge of the base model but also more semantic-related new knowledge so as to beat the base model. We therefore choose a mask-reconstruction (He et al., 2022) SSL scheme for the pretraining of the new model, in which the base model generates reconstruction targets from the full input images and the new model tries to predict base model targets from random masked image input. With this pretext task, the new model is forced to learn the semantic of the full input and also its patch relations so that new model can reason the desired full information from an absent input. As illustrated by Fig. 2, the attentions of iBOT (Zhou et al., 2022a) misses many semantics, *e.g.,* ears, while TEC with iBOT as base model captures all semantics and also well distinguish all different components of an input image. Because of its almost comprehensive semantic capturing ability, TEC can flexibly select related semantics for a downstream task.

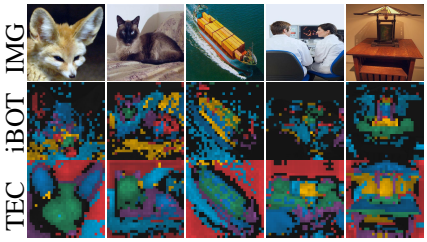

Figure 2: Self-attention visualization. Different colors denote attentions of different heads. Black means no attention.

However, different SSL base models could have various properties due to their various targets and training strategies, *e.g.,* iBOT models with more category semantics while MAE models with more image details (He et al., 2022). So it is important to build high-qualified and compatible reconstruction targets from the base model so that new model learns these targets in a complementary manner. For target quality, a good base model target should reveal the semantic relations among patches, *e.g.,* relation between car wheels and car body, so that new model can learn this general relation patterns and adapts to downstream tasks. To this end, we propose to enhance the target quality of the base models by using two complementary reconstruction targets: a) the patch-dim normalized features which normalizes base model targets along patch dimension to preserve the relations among input patches, and b) semantic attention maps that are patch attention maps with rich semantics to filter out possible background noise and thus to establish the correlation between the whole input semantic and the patch semantic. For target compatibility, we introduce conditional adapters into new model so that new model prediction can be adaptable to various base models with different properties. Given a base model target, the adapters conditionally active and adjust the new model mid-level features to predict the target more effectively. These adapters are discarded after pretraining, but can serve parameter-efficient finetuning (Jia et al., 2022; Chen et al., 2022b) if kept.

We call the above method for sustainable SSL as target-enhanced conditional mask-reconstruction learning (TEC). As shown in Fig. 3, on ImageNet, TEC without any extra training data improves the SSL base model by a remarkable margin, *e.g.,* MAE (He et al., 2022) and iBOT (Zhou et al., 2022a). Moreover, we also find that TEC can significantly accelerate the SSL learning process and saves training cost. For example, training TEC for only 100 epochs with random initialization and a 300-epochs-trained MAE base model outperforms MAE trained with 1600 epochs. We hope our initial effort towards sustainable SSL will inspire more works in the future to sustainably improve SSL in a cost-friendly manner.

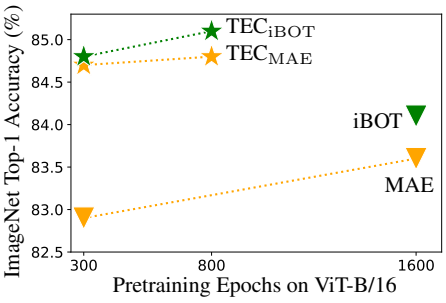

Figure 3: Top1 accuracy on ImageNet-1k.

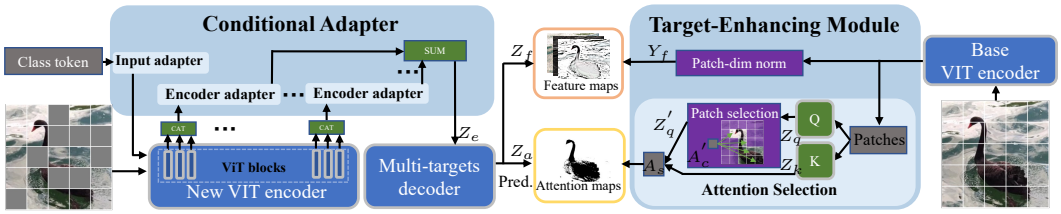

Figure 4: The overall framework of the proposed TEC. The pretrained SSL base model in TEC generates patch-relation enhanced reconstruction targets, *i.e.,* patch-dim normalized features and semantic attention maps. The new ViT encoder takes in masked image patches and class token enhanced by input adapter, and sequentially passes the generated features into encoder adapters and multi-target decoder to predict the base model targets.

2 METHOD

2.1 OVERALL FRAMEWORK

An overall framework of the proposed target-enhanced conditional mask-reconstruction (TEC) method is illustrated in Fig. 4. TEC follows (He et al., 2022; Bao et al., 2021), and uses Vision Transformer (ViT) (Dosovitskiy et al., 2020) for implementation. Under the mask-reconstruction (He et al., 2022) framework, TEC consists of a randomly initialized new ViT encoder to be pretrained, conditional adapters for conditional pretraining, and a multi-target decoder for reconstruction targets prediction, an SSL pretrained ViT encoder as the base model and an target-enhancing module to generate patch-relation enhanced reconstruction base model targets. Specifically, base model ViT encoder is a SSL pretrained encoder (*e.g.,* in MAE (He et al., 2022)), and is used to generate latent representation of a full image. Then target-enhancing module enhances the latent representation to generate two complementary reconstruction targets for new model. The new ViT encoder together with adapters take in masked image and generate adapted latent representation which is then fed into the multi-target decoder to predict the base model targets. After pretraining, the new ViT encoder is kept for downstream tasks while other parts are removed. At below, we explain the conditional pretraining aided by adapters in Section 2.2 to help new model effectively predict base model targets, and then introduce the target-enhancing module to generate high-qualified base model targets in Section 2.3.

2.2 CONDITIONAL PRETRAINING

As aforementioned, different base models often have different properties, *e.g.,* more category semantic in iBOT while more local semantic in MAE. So the prediction of the new model should be compatible to any given base model. To resolve a similar issue on RGB image reconstruction, the works (Wang et al., 2022a; Dong et al., 2022; Gao et al., 2022) manually select certain features from the mid-level layers of the encoder to better align with the RGB image target. However, it is almost impossible to manually select certain fixed encoder features which are compatible to different base models with various properties. Therefore, to improve the learning performance, the new model should have the conditional adaptation ability regarding a given SSL base model.

Given a fixed pretrained model, the parameter-efficient fine-tuning scheme introduces trainable extra modules with a small number of parameters into this pretrained model for adapting it to downstream tasks in both vision (Jia et al., 2022; Chen et al., 2022b) and NLP (Houlsby et al., 2019; Li & Liang, 2021; Liu et al., 2021) domains. For example, the prompting scheme (Li & Liang, 2021; Liu et al., 2021; Jia et al., 2022) concatenates learnable input tokens, *e.g.,* class token, with patch tokens to activate certain semantic features of a fixed ViT model that are suitable for specific downstream task. Also, inserting lightweight adapter modules (*e.g.,* MLP (Houlsby et al., 2019; Chen et al., 2022b) and residual blocks (Li et al., 2022b)) into a fixed model can activate and modulate mid-level features of the model to predict features required by the downstream task. Inspired by these parameter-efficient fine-tuning schemes, we apply the adaptation scheme in the pretraining stage by introducing conditional adapters into the new model to handle the diversities of base models. Our adapters are only used for pretraining and will be removed during finetuning. So they do not increase extra inference cost. Actually, Tab. 4 shows that keeping these adapters in the inference phase would

enhance the parameter-efficient finetuning ability of the model. At below, we will introduce how to apply adapters, *i.e.,* input and encoder adapters, into the new model.

Input adapter. For ViT networks, one often concatenate a class token with the input patch tokens to learn the global semantic of the whole input. Since the prompting scheme shows the adaption ability of the class token, we propose to further enhance the feature adaption ability of class token by adding an input adapter. The input adapter, composed of a small two-layer MLP layer, as shown in Fig. 5, enhances the representation ability the class token so that the class token can better activate features in the new model according to the base model targets supervision. Specifically, the class token $T \in \mathbb{R}^C$ of the ViT is processed by the MLP layer to obtain an enhanced class token $T^{'} \in \mathbb{R}^C$:

$$T^{'} = \mathbf{MLP}(T),$$

where C is the embedding dimension. During pretraining, $\mathbf{MLP}(T)$ is appended to the patch tokens. $\mathbf{MLP}$ enhances the representation ability of T and enables the new model to better predict base model targets with stronger adaptability. $\mathbf{MLP}(T)$ is merged into a class token $T^{'}$ by applying the reparameterization trick (Ding et al., 2021) for inference without extra inference cost.

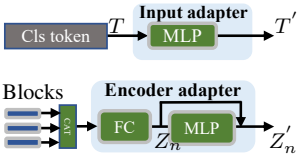

Figure 5: The input adapter and encoder adapter.

Encoder adapter. To modulate mid-level features in the new model so that it can adapt to the base model targets, we apply a simple MLP with residual connection (Chen et al., 2022b) as our encoder adapter in the pretraining phase. As we want to remove adapters after pretraining to save inference cost, we need to keep the encoder network topology unchanged after removing adapters. So we put the input of adapters in the middle of the encoder and merge all adapter outputs at the end of the encoder. As shown in Fig. 5, given features $X = \{X_i, i = 1, \ldots, D\}$ from each block of the new model encoder where D is the number of encoder blocks, we first uniformly divide them into N groups, in which each group contains 3 blocks by default. Within the n_{th} group, we merge features from adjacent blocks to obtain feature Z_n which can save computational cost:

$$Z_n = \mathbf{FC}(\mathbf{Concat}(X_i, ..., X_j)).$$

Then we further feed the feature Z_n into an adapter and obtain feature Z_e:

$$Z_n^{'} = Z_n + \mathbf{MLP}(Z_n), \qquad Z_e = \sum\nolimits_{n=1}^{N} Z_n^{'}, \tag{1}$$

where is $\mathbf{MLP}$ is small MLP with 2 layers. The adapted features are then sent to the multi-target decoder to predict base model targets, which will be introduced in Section 2.3.

2.3 Patch-relation enhanced Reconstruction Targets

To better exploit the knowledge of base models in sustainable SSL, in the target-enhancing module, we construct two complementary targets with enhanced patch-relations: 1) feature-level targets with patch dimension normalization to strengthen the relations between patches; 2) semantic-related attention maps to learn relations between semantic-related patches and other patches. The feature targets reveal semantics of certain patches, while attention maps focus more on feature relations.

Patch-dim normalized feature-level targets. Given a base model, we propose to normalize its target along the patch dimension to preserve the spatial patch-relations. Specifically, for an input, assume its base model target is $Y \in \mathbb{R}^{L \times C}$ where L and C respectively denote patch number and channel dimension C. Then we normalize Y along patch dimension:

$$Y_f = (Y - \mu_L)/\sigma_L, \tag{2}$$

where μ_L and σ_L are mean and variance along the patch dimension. For MIM, this patch-dim normalization can better preserve the spatial relations among tokens than the widely used feature normalization (Wei et al., 2022c;a; Baevski et al., 2022) along channel dimension. This is because the base model features might have large variances, *e.g.,* a few values are much larger than other values. Normalization along channel dimension would lead to a degenerated case that only a few feature values are large while other features have very small values. Thus, patches within one channel might have very similar values, impair their spatial relations. In contrast, patch-dim normalization would relieve this issue, and still preserve the possible inherent spatial relations among patches. Indeed, Tab. 7(a) shows that such simple modification significantly improves the new model performance.

After normalization, following (He et al., 2022), new model uses an fully-connected layer followed by the decoder to generate Z_f for predicting the base model target Y_f on masked patches:

$$L_{\text{fea}} = \|M \circ (Y_f - Z_f)\|_2^2, \tag{3}$$

where M is the mask matrix and $\circ$ denotes the element-wise product.

Semantic attention-level targets. Self-attention in pretrained ViT models has a powerful capability of capturing semantic relations among patch tokens (Caron et al., 2021; Li et al., 2022c). We then propose to utilize the self-attention maps as a type of reconstruction targets for MIM to further enhance the semantic relation modeling capability of the new model. According to previous investigations on effects of self-attention maps in KD (Wu et al., 2022; Wang et al., 2022b), not all attention maps contain useful semantic relations, and severe noisy attentions even hinder student learning. Accordingly, it is necessary to select parts of attention maps for reducing the possible severe noise and also helping reduce training cost. Here we utilize the base model class token which contains sufficient global semantics to select top similar patch tokens which filters out the possible noises. Specifically, given the attention maps $A_c \in \mathbb{R}^{H\times L}$ between class token and patch tokens from the last ViT block in base model where L and H respectively denote patch number and head number, we average the attention map A_c along head dimension to obtain $A_c^{'} \in \mathbb{R}^{1\times L}$. Then, as shown in Fig. 4, we select top-k patches with the largest values in $A_c^{'}$, and further compute the attention map $A_p \in \mathbb{R}^{H\times k\times L}$ among the top-k patches and all patch tokens. Consider the importance of class token, we further concatenate attention maps between itself and selected A_p to obtain our final reconstruction targets, *i.e.,* $A_s \in \mathbb{R}^{H\times(k+1)\times L}$. Note, when we compute A_s, a temperature τ is added before the Softmax operation to adjust the attention sharpness. For the new model, we respectively use two fully-connected layers to project its decoder output into two predictions $Z_q \in \mathbb{R}^{L\times C}$ and $Z_k \in \mathbb{R}^{L\times C}$. We select the same patches as in A_s from Z_q to form $Z_q^{'} \in \mathbb{R}^{(k+1)\times C}$. Then we concatenate the class token cls in new model with $Z_q^{'}$ and compute the KQ attention map $Z_a = \text{Softmax}([Z_q^{'}, \mathsf{cls}]^\top Z_k) \in \mathbb{R}^{H\times(k+1)\times L}$. Finally, we compute the prediction loss between the prediction Z_a and the teacher target A_s via the entropy loss:

$$L_{\text{att}} = -A_s \log Z_a. \tag{4}$$

Multi-target decoder. Due to the different property of the two reconstruction targets, namely feature target and attention target, one decoder in new model for prediction is not sufficient to handle them at the same time and often tends to results in prediction conflict. But using separate decoders for each target would increase the trainable parameters and thus training cost. To solve this problem, we use a simple decoder adaptation scheme which constructs target-specific inputs and then feed them into a shared decoder. Specifically, we feed the output feature Z_e (see Eqn. 1) of new model encoder into a fully-connected layer and then fill the masked tokens by a learnable mask token to obtain $Z_f^{'}$. Then similarly, given Z_e, we can also use a fully-connected layer and a learnable mask token to obtain $Z_m^{'}$. Next, we respectively feed $Z_f^{'}$ and $Z_m^{'}$ into a shared transformer-based decoder for predicting the feature and attention map targets of base model. Unlike the large semantic gap between encoder output and vanilla image in MAE, the base model target have similar semantics to the new model predictions. So a shallow 2-layer decoder is enough and is indeed better than the 8-layer decoder used in MAE. This also greatly reduces the training cost.

3 Experiments

We evaluate our TEC on ImageNet-1k (Deng et al., 2009) and use it to train ViT (Dosovitskiy et al., 2020) with a 16×16 patch size and 224×224 image resolution. For all experiments, we randomly initialize the models and pretrain for 300/800 epochs via AdamW (Loshchilov & Hutter, 2017) of 4,096 batchsize (He et al., 2022). To make sure the gain is from our sustainable SSL method instead of extra data or stronger base model , explicit/implicit extra training data or stronger base models than new models (Ramesh et al., 2021) are not used. So we use iBOT (Zhou et al., 2022a) and MAE (He et al., 2022) pretrained ViT models on ImageNet-1k as our base models. We apply the same masking strategy as in MAE, *e.g.,* 75% masked ratio. See more training details in Appendix.

3.1 Performance comparison

3.1.1 Comparison on the ImageNet dataset

Fully finetuning on ImageNet-1k. We compare the fully finetuning performance of ViT-B on ImageNet-1k in Tab. 1. It is observed that our TEC outperforms the iBOT base model with 0.7%

Table 1: Comparison with existing SSL methods under ImageNet-1k fully finetuning using ViT. †and gray color mean trained with implicit /explicit extra data.

Model	Method	Epoch	Guidance	Top1 acc.
ViT-Base	DINO (Caron et al., 2021)	300	NA	82.8
	MoCov3 (Chen* et al., 2021)	300	NA	83.2
	MixMIM (Liu et al., 2022b)	300	RGB	83.2
	MFM (Xie et al., 2022a)	300	Frequency	83.1
	BEiT (Bao et al., 2021)	800	DALLE†	83.2
	SplitMask (El-Nouby et al., 2021)	300	NA	83.6
	ConMIM (Yi et al., 2022)	800	Momentum	83.7
	SimMIM (Xie et al., 2022b)	800	RGB	83.8
	SIM (Tao et al., 2022)	1600	Momentum	83.8
	CAE (Chen et al., 2022c)	1600	DALLE†	83.9
	MaskFeat (Wei et al., 2022a)	1600	HOG	84.0
	LoMaR (Chen et al., 2022a)	1600	RGB	84.1
	BootMAE (Dong et al., 2022)	800	RGB+Momentum	84.2
	data2vec (Baevski et al., 2022)	800	Momentum	84.2
	Mugs (Zhou et al., 2022b)	1600	NA	84.3
	MVP (Wei et al., 2022b)	300	CLIP†	84.4
	PeCo (Dong et al., 2021)	800	Perceptual codebook	84.5
	CMAE (Huang et al., 2022)	1600	RGB	84.7
	Ge2-AE (Liu et al., 2022a)	800	RGB+Frequency	84.8
	FD-CLIP (Wei et al., 2022c)	300	CLIP†	84.9
	MAE (He et al., 2022)	1600	RGB	83.6
	FD-MAE (Wei et al., 2022c)	300	MAE	$83.8_{+0.2}$
	TEC	300	MAE	$\mathbf{84.7}_{+1.1}$
	TEC	800	MAE	$\mathbf{84.8}_{+1.2}$
	iBOT-ImageNet-22K	-	Momentum	84.4
	iBOT (Zhou et al., 2022a)	1600	Momentum	84.1
	SemMAE (Li et al., 2022a)	800	iBOT	$84.5_{+0.4}$
	TEC	300	iBOT	$\mathbf{84.8}_{+0.7}$
	TEC	800	iBOT	$\mathbf{85.1}_{+1.0}$
ViT-Large	MAE (He et al., 2022)	1600	RGB	85.9
	TEC	300	MAE	$\mathbf{86.5}_{+0.6}$

Table 2: Semantic segmentation on ADE20k using Upernet and ViT-B.

Method	Epoch	mIoU
BEiT	800	47.1
PeCo	800	48.5
GE2-AE	800	48.9
CAE	1600	50.2
CMAE	1600	50.1
MAE	1600	48.1
$\mathbf{TEC}_{\mathrm{MAE}}$	800	**49.9**
iBOT	1600	50.0
$\mathbf{TEC}_{\mathrm{iBOT}}$	800	**51.0**

Table 3: Instance segmentation on COCO using Cascade MaskRCNN and ViT-B.

Method	$\mathrm{AP}_{\mathrm{bbox}}$	$\mathrm{AP}_{\mathrm{mask}}$
Implementation from (Zhou et al., 2022a)		
iBOT	51.2	44.2
$\mathbf{TEC}_{\mathrm{iBOT}}$	**52.7**	**45.4**
Implementation from (Li et al., 2022b)		
MAE	54.0	46.7
$\mathbf{TEC}_{\mathrm{MAE}}$	**54.4**	**47.1**

when trained with 300 epochs from random initialization. When trained with 800 epochs, the gain is enlarged to 1.0%, showing TEC benefits from longer training. Similarly, TEC brings gains of 1.1% and 1.2% over the MAE base model with 300/800 epochs pretraining. Though MAE and iBOT are both strong MIM-based methods, TEC can still further improve them with the proposed target-enhanced conditional MIM scheme, verifying its effectiveness. With the help of SSL base models, TEC outperforms other SSL methods with similar or less training cost, including methods trained with implicitly extra data, *e.g.,* MVP (Wei et al., 2022b) and FD-CLIP (Wei et al., 2022c). Compared to iBOT trained with large-scale ImageNet-22k dataset, TEC has the gain of 0.7% with only ImageNet-1k data, indicating TEC is more effective than enlarging the training cost of the base model. To the best of our knowledge, TEC achieves new SOTA on ViT-B when solely using ImageNet-1k training data, showing the potential of sustainable SSL learning. We also verify the scaling ability of our TEC by using the ViT-L model, and find that TEC surpasses the MAE base model by 0.6% with 300 epochs pretraining from random initialization.

Parameter-efficient finetuning on ImageNet-1k. Parameter-efficient finetuning methods, *e.g.,* linear probing, enable finetuning a small number of parameters on the downstream task. Here we test effectiveness of our TEC with parameter-efficient finetuning for classification on ImageNet-1k using ViT-B, with results shown in Tab. 4. Under the linear probing setting, TEC outperforms the MAE

Table 4: Top1 accuracy on the ImageNet-1k dataset under parameter-efficient finetuning.

Method	Epoch	Settings	Top 1 acc.
MAE	1600	Linear probing	68.0
TEC_{MAE}	800	Linear probing	69.8
		+Input adapter FT	72.6
		+Encoder adapter FT	79.9

Table 5: Semi-supervised semantic segmentation on the ImageNet-S dataset.

Pretrain	Method	Epoch	mIoU_{val}
SSL	MAE	1600	38.3
	TEC_{MAE}	800	**42.9**
SSL+FT	MAE	1600+100	61.0
	TEC_{MAE}	800+100	**62.0**

base model with 1.8%, indicating the learnt new model contains more category-related semantic information. The input and encoder adapters for pretraining can also be used for parameter-efficient finetuning. Finetuning with the input adapter, *i.e.,* prompting, brings a considerable gain of 4.6%. When finetuned with input and encoder adapters, the performance gain with MAE is enlarged to 11.9%. Therefore, adapters used for pretraining can greatly benefit parameter-efficient finetuning.

Semantic segmentation on ImageNet-S. To test the pixel-level representation ability of TEC pretrained models, we conduct semantic segmentation finetuning on ImageNet-S (Gao et al., 2021) that has pixel-level training labels. We use ViT-B as the segmentation model without extra segmentation head, since the pretraining and finetuning data have no domain shift. Tab. 5 shows that when finetuning with SSL pretrained models, TEC_{MAE} improves MAE base model by 4.6% mIoU. When using supervised ImageNet fully-finetuned pretraining, TEC_{MAE} achieves a gain of 1.0% over MAE.

3.1.2 Transfer learning on downstream tasks

Here we investigate the transfer learning ability of TEC models on downstream tasks.

Semantic segmentation. For semantic segmentation on the ADE20k (Zhou et al., 2018) dataset, we use Upernet (Xiao et al., 2018) with ViT-B as the segmentation model. Tab. 2 shows TEC_{iBOT} surpasses the iBOT base model by 1.0% mIoU, and TEC_{MAE} achieves a 1.8% gain over its MAE base model. It can be seen that TEC pretrained models show greater transfer learning abilities on semantic segmentation compared to their base models. Also, TEC shows clear advantages over strong competitors with fewer pretraining epochs. For example, it outperforms MAE, CAE (Chen et al., 2022c), and CMAE (Huang et al., 2022) by 2.9%, 0.8% and 0.9%, achieving new SOTA.

Instance segmentation. For instance segmentation on the COCO (Lin et al., 2014) dataset, to make fair comparisons and save reproduce costs, we apply the Cascade MaskRCNN (Cai & Vasconcelos, 2019) implemented by iBOT (Zhou et al., 2022a) and ViTDet (Li et al., 2022b) for TEC with iBOT/MAE base models. Tab. 3 shows that TEC surpasses the iBOT base model by 1.5% on box AP and 1.2% on mask AP. Using the Cascade MaskRCNN implementation from ViTDet, TEC still achieves a gain of 0.4% on box AP and 0.4% on mask AP, indicating stable improvements of TEC.

3.2 Ablation and Analysis

We give the ablation study and analysis of our proposed method. By default, models are pretrained with 300 epochs and evaluated with the fully finetuning on ImageNet-1k.

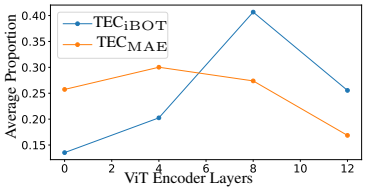

Figure 6: The average proportion of encoder adapters contributing to the encoder output Z_e.

Conditional pretraining. The conditional modules, *i.e.,* input/encoder adapters, aid the SSL pretraining under different base models. Tab. 8 shows adapters stably improve the performance by 0.4% and 0.2% when using MAE and iBOT as base models. In Tab. 7(b), with MAE base model, input adapters improve 0.1% over baseline and encoder adapters further brings a gain of 0.3%. To observe the adaptation difference to base models, we show the average proportion of encoder adapters contributing to the encoder output Z_e in Fig. 6. iBOT base model requires adapters to provide more features from deeper layers, while MAE base model makes adapters focus more on shallow layers, which is constant with their properties, *i.e.,* iBOT base model has more high-level category semantics while MAE model has more low-level image details.

Feature normalization on different dimensions. We normalize target features on the patch dimension to stress the relative relations among patches, which differs from existing methods that normalize on the channel dimension. In Tab. 7(a), normalizing on patch dimension achieves 0.3% gain than channel-dim normalization. While the channel-dim normalization has no effect compared to the unnormalized version. Channel-dim normalization emphasizes the feature difference in channels. Instead, our patch-dim normalization stresses the relations among patches, which is compatible

Table 7: Ablation study on ImageNet-1K fully finetuning setting using ViT-B.

(a) Patch-norm features.

	Top1 acc.
MAE base	83.6
NA	83.9
Feature dim.	83.9
Patch dim.	**84.2**

(b) Effect of adapters.

	Top1 acc.
MAE base	83.6
No adapter	84.2
+ input adapter	84.3
+ encoder adapter	84.6

(c) Init. with base model pretrain.

	Top1 acc.
iBOT base	84.1
Load	84.4
Not load	**84.8**

(d) TEC accelerates MAE training.

	Epoch	Top1 acc.
MAE	1600	83.6
TEC$_{\text{MAE1600ep}}$	300	**84.7**$_{+1.1}$
MAE	300	82.9
TEC$_{\text{MAE300ep}}$	**100**	**83.9**$_{+1.0}$
TEC$_{\text{MAE300ep}}$	300	**84.3**$_{+1.4}$

(e) Effect of semantic-related patch attention.

	Top1 acc.
iBOT base	84.1
No attention	84.5
Cls token only	84.5
All attention	84.6
Attention select	**84.7**

Table 8: Ablation study on ImageNet-1K fully finetuning setting using ViT-B.

Patch-norm. feature	Semantic attention	Adapters	MAE base	iBOT base
Base model performance			83.6%	84.1%
✓			84.2%	84.5%
✓	✓		84.3%	84.7%
✓		✓	84.6%	84.7%
✓	✓	✓	84.7%	84.8%

to the patch-prediction in the MIM scheme. Tab. 8 shows training with patch-dim normalized feature has the 0.6%/0.4% gain over MAE/iBOT base models, showing its robustness over base models.

Semantic-related attention. The KQ attention maps naturally contain the semantic relations among patches, thus are used as the targets to enhance the patch-relation property of base model targets. Tab. 8 shows that using attention maps further improves the models trained with patch-dim normalized features. Tab. 7(e) compares the effects of different types of attention maps. Only using the attention maps of the class token has no improvement, while the attention of semantic-related patches brings 0.2% gain over baseline. Therefore, it is the relation among patches that helps the MIM training. Compared to using all attention maps, using the selected semantic-related attention maps brings larger gain by reducing the noise.

Accelerating the training process of base models. By default, we use the fully pretrained SSL models as base models. To verify if the proposed TEC can improve an unconverged SSL model, we use a 300-epoch MAE pretrained ViT-B model as the base model and train TEC with 100/300 epoch from random initialization. As shown in Tab. 7(d), the 300-epoch pretrained MAE gives the performance of 82.9%. In comparison, TEC$_{\text{MAE300ep}}$ achieves 84.3%/83.9% with 300/100 epochs, surpassing the 300-epoch MAE base models with 1.4%/1.0%. Notably, TEC$_{\text{MAE300ep}}$ even outperforms the 1600-epoch pretrained MAE by 0.3% with only 100 epoch training, showing TEC can significantly accelerate the training process of the base model. Still, the TEC$_{\text{MAE1600ep}}$ taught with 1600-epoch MAE base model further improves the TEC$_{\text{MAE300ep}}$ by 0.4%, indicating our sustainable learning scheme relies on good base models to achieve better performance.

Towards general sustainable SSL. To make more steps towards sustainable SSL, we use the TEC pretrained models as the base model for a new round of TEC pretraining. Tab. 6 shows that the second-round TEC trained with the first-round TEC base model achieves 85.2%. The possible reason for the smaller improvement in the second round is caused by the limited network capacity or two-rounds of TEC pretraining learns similar knowledge.

Table 6: Towards general sustainable SSL by using the TEC as new base model.

Model	Base	Epoch	Top1 acc.
iBOT	-	1600	84.1
TEC$_{\text{iBOT}}$	iBOT	800	85.1
TEC	TEC$_{\text{iBOT}}$	800	85.2

Initializing new model with base model weights or not. The new models in the TEC framework are trained from random initialization. Tab. 7(c) compares the new model performance with/without loading the pretrained weights of base models. Randomly initialized new model outperforms the model loaded with pretrained base model weights by 0.4%. We assume that randomly initialized new models avoid the local minima of the base model, and the new model learns a different weight distribution compared to the base model.

4 RELATED WORKS

Self-supervised learning. Self-supervised learning enables the representation pretraining without human annotation. Instance discrimination task (ID), *e.g.,* (He et al., 2020; Zhou et al., 2022b), and masked image modeling (MIM), *e.g.,* (He et al., 2020), are two mainstream pipelines. Instance discrimination task, *e.g.,* contrastive-based (Chen et al., 2020; He et al., 2020), non-contrastive-based (Grill et al., 2020; Chen & He, 2021; Zbontar et al., 2021; Li et al., 2022c), and clustering-based (Caron et al., 2018; 2020; 2021), conduct learning by pulling close representations from multiple views of one image. Benefited from this property, discriminative methods learn high category-related representations. Originally designed for NLP (Devlin et al., 2018), MIM pushes the vision SSL to new SOTA with good scaling ability to large models, showing great potential on image representation learning (Bao et al., 2021; Zhou et al., 2022a; Xie et al., 2022b). MIM learns semantics by reconstructing the information of masked patches from unmasked parts, which learns more spatial semantic details than ID. (Huang et al., 2022; Wang et al., 2022a) explore combining the advantages of MIM and ID to further improve performance. Recently, (Kong & Zhang, 2022) reveals both MIM and ID learn occlusion-invariant features. We observe a trend that these SSL methods require increasingly large computing cost to achieve SOTA, which hinders the development of new SSL methods. To remedy this, we explore sustainable SSL by learning from pretrained SSL models.

Masked image modeling on various targets. The reconstruction targets guide the MIM learning on different semantic spaces. MIM has explored various reconstruction targets, *e.g.,* RGB pixels and tokenizers. To make images similar to the discretized language in NLP (Devlin et al., 2018), Beit (Bao et al., 2021) utilizes the DALLE pretrained tokenizer (Ramesh et al., 2021) as the prediction target. CAE (Chen et al., 2022c) further decouples the pretext task prediction with the encoder. MAE and SimMIM (Xie et al., 2022b) show using RGB images as targets achieves competitive fully finetuning performance. MaskFeat (Wei et al., 2022a) reveals hand-designed HOG feature (Dalal & Triggs, 2005) is an effective target form. Ge2-AE (Liu et al., 2022a) and MFM (Xie et al., 2022a) find the image frequency domain can be complementary to RGB image targets. The perceptual codebook in PeCo (Dong et al., 2021) helps the model learn semantic information. The online momentum network (He et al., 2020) is used by iBOT and data2vec (Baevski et al., 2022) to provide updated prediction targets. BootMAE (Dong et al., 2022) takes the advantage of both RGB image and online network targets. (Yang et al., 2022) enhances the distillation from a large teacher model to a compact student model with the masking scheme. MVP (Wei et al., 2022b) introduces rich semantics learned from vision-language pretraining by using the CLIP pretrained model (Ramesh et al., 2021) as targets. Unlike these works that stress the unique properties of a specific reconstruction target, we show that all SSL pretrained models can serve as good base models with the help of our TEC. Our method achieves better performance with the assistance of existing SSL models, making the cost-friendly sustainable SSL learning possible.

Self-supervised knowledge distillation. The self-supervised knowledge distillation is similar to sustainable learning in this work as they both learn from SSL pretrained models. However, the knowledge distillation methods mostly aim to transfer knowledge from strong and large models to compact new models. SEED (Fang et al., 2021) distillates knowledge from large SSL models to small models with contrastive loss. (Navaneet et al., 2022) uses MLP heads for feature regression to distill large SSL teachers to compact models. (Xu et al., 2021) groups the instances with teacher models and transfers the instance-relation knowledge to new models. As an exception, (Wei et al., 2022c) shows feature distillation improves contrastive-based SSL models but brings marginal gain over the SOTA MAE (He et al., 2022) model. In contrast, our method for sustainable learning lets the new model surpass the base model with the same model size and training data.

5 CONCLUSION

This work explores sustainable self-supervised learning by learning from pretrained SSL models. We propose a target-enhanced conditional mask-reconstruction learning scheme to learn from and surpass existing SSL models. The adapters help to adapt the new model to various base models during pretraining and can also serve as parameter-efficient finetuning modules. We utilize the mask-reconstruction scheme as the basis for surpassing base models, and we enhance the patch-level relation of prediction targets to aid the MIM pretraining. Our method further improves the strong MIM pretrained methods, *e.g.,* MAE and iBOT, proving feasibility of sustainable learning.

REFERENCES

Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. Data2vec: A general framework for self-supervised learning in speech, vision and language. *arXiv preprint arXiv:2202.03555*, 2022.

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

Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: high quality object detection and instance segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 43(5): 1483–1498, 2019.

Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *European Conference on Computer Vision (ECCV)*, 2018.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *IEEE International Conference on Computer Vision (ICCV)*, 2021.

Jun Chen, Ming Hu, Boyang Li, and Mohamed Elhoseiny. Efficient self-supervised vision pretraining with local masked reconstruction. *arXiv preprint arXiv:2206.00790*, 2022a.

Shoufa Chen, Chongjian Ge, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. Adaptformer: Adapting vision transformers for scalable visual recognition. *arXiv preprint arXiv:2205.13535*, 2022b.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning (ICML)*, pp. 1597–1607. PMLR, 2020.

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*, 2022c.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021.

Xinlei Chen*, Saining Xie*, and Kaiming He. An empirical study of training self-supervised vision transformers. *arXiv preprint arXiv:2104.02057*, 2021.

MMSegmentation Contributors. MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. https://github.com/open-mmlab/mmsegmentation, 2020.

Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pp. 886–893. Ieee, 2005.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 248–255. Ieee, 2009.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Xiaohan Ding, Xiangyu Zhang, Ningning Ma, Jungong Han, Guiguang Ding, and Jian Sun. Repvgg: Making vgg-style convnets great again. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13733–13742, 2021.

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

Xiaoyi Dong, Jianmin Bao, Ting Zhang, Dongdong Chen, Weiming Zhang, Lu Yuan, Dong Chen, Fang Wen, and Nenghai Yu. Bootstrapped masked autoencoders for vision bert pretraining. In *European Conference on Computer Vision (ECCV)*, 2022.

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

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

Zhiyuan Fang, Jianfeng Wang, Lijuan Wang, Lei Zhang, Yezhou Yang, and Zicheng Liu. Seed: Self-supervised distillation for visual representation. *arXiv preprint arXiv:2101.04731*, 2021.

Peng Gao, Teli Ma, Hongsheng Li, Jifeng Dai, and Yu Qiao. Convmae: Masked convolution meets masked autoencoders. *arXiv preprint arXiv:2205.03892*, 2022.

Shanghua Gao, Zhong-Yu Li, Ming-Hsuan Yang, Ming-Ming Cheng, Junwei Han, and Philip Torr. Large-scale unsupervised semantic segmentation. *arXiv preprint arXiv:2106.03149*, 2021.

Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International Journal of Computer Vision (IJCV)*, 129(6):1789–1819, 2021.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Ávila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent - a new approach to self-supervised learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

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

Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7), 2015.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning (ICML)*, pp. 2790–2799. PMLR, 2019.

Zhicheng Huang, Xiaojie Jin, Chengze Lu, Qibin Hou, Ming-Ming Cheng, Dongmei Fu, Xiaohui Shen, and Jiashi Feng. Contrastive masked autoencoders are stronger vision learners. *arXiv preprint arXiv:2207.13532*, 2022.

Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. *arXiv preprint arXiv:2203.12119*, 2022.

Xiangwen Kong and Xiangyu Zhang. Understanding masked image modeling via learning occlusion invariant feature. *arXiv preprint arXiv:2208.04164*, 2022.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International Conference on Machine Learning (ICML)*, pp. 3519–3529. PMLR, 2019.

Gang Li, Heliang Zheng, Daqing Liu, Bing Su, and Changwen Zheng. Semmae: Semantic-guided masking for learning masked autoencoders. *arXiv preprint arXiv:2206.10207*, 2022a.

Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*, 2021.

Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. *arXiv preprint arXiv:2203.16527*, 2022b.

Zhong-Yu Li, Shanghua Gao, and Ming-Ming Cheng. Exploring feature self-relation for self-supervised transformer. *arXiv preprint arXiv:2206.05184*, 2022c.

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 *European Conference on Computer Vision (ECCV)*, pp. 740–755, 2014.

Hao Liu, Xinghua Jiang, Xin Li, Antai Guo, Deqiang Jiang, and Bo Ren. The devil is in the frequency: Geminated gestalt autoencoder for self-supervised visual pre-training. *arXiv preprint arXiv:2204.08227*, 2022a.

Jihao Liu, Xin Huang, Yu Liu, and Hongsheng Li. Mixmim: Mixed and masked image modeling for efficient visual representation learning. *arXiv preprint arXiv:2205.13137*, 2022b.

Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. Gpt understands, too. *arXiv preprint arXiv:2103.10385*, 2021.

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

KL Navaneet, Soroush Abbasi Koohpayegani, Ajinkya Tejankar, and Hamed Pirsiavash. Simreg: Regression as a simple yet effective tool for self-supervised knowledge distillation. *arXiv preprint arXiv:2201.05131*, 2022.

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 (ICML)*, pp. 8821–8831. PMLR, 2021.

Chenxin Tao, Xizhou Zhu, Gao Huang, Yu Qiao, Xiaogang Wang, and Jifeng Dai. Siamese image modeling for self-supervised vision representation learning. *arXiv preprint arXiv:2206.01204*, 2022.

Luya Wang, Feng Liang, Yangguang Li, Wanli Ouyang, Honggang Zhang, and Jing Shao. Repre: Improving self-supervised vision transformer with reconstructive pre-training. *arXiv preprint arXiv:2201.06857*, 2022a.

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*, 2022b.

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

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

Yixuan Wei, Han Hu, Zhenda Xie, Zheng Zhang, Yue Cao, Jianmin Bao, Dong Chen, and Baining Guo. Contrastive learning rivals masked image modeling in fine-tuning via feature distillation. *arXiv preprint arXiv:2205.14141*, 2022c.

Haiyan Wu, Yuting Gao, Yinqi Zhang, Shaohui Lin, Yuan Xie, Xing Sun, and Ke Li. Self-supervised models are good teaching assistants for vision transformers. In *International Conference on Machine Learning (ICML)*, pp. 24031–24042. PMLR, 2022.

Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *European Conference on Computer Vision (ECCV)*, pp. 418–434, 2018.

Jiahao Xie, Wei Li, Xiaohang Zhan, Ziwei Liu, Yew Soon Ong, and Chen Change Loy. Masked frequency modeling for self-supervised visual pre-training. *arXiv preprint arXiv:2206.07706*, 2022a.

Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10687–10698, 2020.

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

Haohang Xu, Jiemin Fang, Xiaopeng Zhang, Lingxi Xie, Xinggang Wang, Wenrui Dai, Hongkai Xiong, and Qi Tian. Bag of instances aggregation boosts self-supervised distillation. In *International Conference on Learning Representations (ICLR)*, 2021.

I Zeki Yalniz, Hervé Jégou, Kan Chen, Manohar Paluri, and Dhruv Mahajan. Billion-scale semi-supervised learning for image classification. *arXiv preprint arXiv:1905.00546*, 2019.

Zhendong Yang, Zhe Li, Mingqi Shao, Dachuan Shi, Zehuan Yuan, and Chun Yuan. Masked generative distillation. *arXiv preprint arXiv:2205.01529*, 2022.

Kun Yi, Yixiao Ge, Xiaotong Li, Shusheng Yang, Dian Li, Jianping Wu, Ying Shan, and Xiaohu Qie. Masked image modeling with denoising contrast. *arXiv preprint arXiv:2205.09616*, 2022.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. *CoRR*, abs/2103.03230, 2021.

Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision (IJCV)*, 2018.

Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. *International Conference on Learning Representations (ICLR)*, 2022a.

Pan Zhou, Yichen Zhou, Chenyang Si, Weihao Yu, Teck Khim Ng, and Shuicheng Yan. Mugs: A multi-granular self-supervised learning framework. In *arXiv preprint arXiv:2203.14415*, 2022b.

A APPENDIX

Pretraining settings on ImageNet-1k. We follow the standard ViT network implemented in MAE. We give the pretraining settings in Tab. 9, which follows the training settings in MAE. Due to different properties of SSL teacher models, we set different parameters of semantic-related attention for MAE and iBOT teacher models as shown in Tab. 10.

Inspired by (Baevski et al., 2022), we utilize the average of the last 2 blocks of the teacher model as the target. We measure the CKA similarity (Kornblith et al., 2019) of each mid-layer block to the output of the last block, and we observe the high feature similarity of last two blocks. Therefore, average features of last 2 layers are used.

Fully finetuning settings on ImageNet-1k. We give the fully finetuning settings on ImageNet-1k in Tab. 11. We observe that TECs trained with different teachers may have different properties. Since their teachers have different layer decay values, we set different layer decay values for TECs trained with different teachers.

Parameter-efficient finetuning settings on ImageNet-1k. Following MAE, we set linear probing settings are shown in Tab. 12. For parameter-efficient finetuning with the input adapter, we use the same training settings as used by the liner probing in Tab. 12. When finetuning with the encoder adapters, we use the same training settings as used by the fully finetuning in Tab. 11 due to more parameters are contained in encoder adapters.

Semi-supervised semantic segmentation finetuning on ImageNet-S. We give the training settings of semi-supervised semantic segmentation finetuning on ImageNet-S in Tab. 13. We set different learning rates and layer decay weights for initializing with pretrained weights with/without fully finetuning.

Downstream task settings. For semantic segmentation on ADE20K, we use the MMSegmentation (Contributors, 2020) implementation of Upernet. The training configurations are following the MAE training configuration in MMSegmentation. For instance segmentation on COCO, we follow the training configurations of iBOT and ViTDet for TEC_{iBOT} and TEC_{MAE} with no change.

Table 9: Pretraining settings.

Configuration	Value
Optimizer	AdamW
Base learning rate	1.5e-4
Weight decay	0.05
Optimizer momentum	β_1, β_2=0.9, 0.95
Batch size	4096
Learning rate schedule	Cosine decay
Warmup epochs	40
Augmentation	RandomResizedCrop

Table 10: Parameters of semantic-related attention.

Teacher	τ	k
MAE (ViT-base)	1.8	15
MAE (ViT-large)	1.4	15
iBOT (ViT-base)	1.0	9

Table 11: Settings of fully finetuning and parameter-efficient finetuning with encoder adapters.

Configuration	Value
Optimizer	AdamW
Base learning rate	1e-3
Min learning rate	1e-6 (B), 1e-5(L)
Weight decay	0.05
Optimizer momentum	β_1, β_2=0.9, 0.999
Layer-wise lr decay	0.55 (MAE-B), 0.65 (iBOT-B), 0.65 (MAE-L)
Batch size	1024
Learning rate schedule	Cosine decay
Warmup epochs	20 (B), 5 (L)
Training epochs	100 (B), 50 (L)
Augmentation	RandAug (9, 0.5)
Label smoothing	0.1
Mixup	0.8
Cutmix	1.0
Drop path	0.1

Table 12: Settings of linear probing and parameter-efficient finetuning with the input adapter.

Configuration	Value
Optimizer	LARS
Base learning rate	0.1
Weight decay	0
Optimizer momentum	0.9
Batch size	16384
Learning rate schedule	Cosine decay
Warmup epochs	10
Training epochs	90
Augmentation	RandomResizedCrop

Table 13: Settings of semantic segmentation finetuning on ImageNet-S.

Configuration	Value
Optimizer	AdamW
Base learning rate	5e-4 (SSL), 1e-4 (SSL+FT)
Weight decay	0.05
Optimizer momentum	β_1, β_2=0.9, 0.999
Layer-wise lr decay	0.60 (SSL), 0.45 (SSL+FT)
Batch size	256
Learning rate schedule	Cosine decay
Warmup epochs	5
Training epochs	100
Augmentation	RandomResizedCrop
Drop path	0.1