The Dynamic Duo of Collaborative Masking and Target for Advanced Masked Autoencoder Learning

Shentong Mo

Carnegie Mellon University

Abstract

Masked autoencoders (MAE) have recently succeeded in self-supervised vision representation learning. Previous work mainly applied custom-designed (*e.g.*, random, block-wise) masking or teacher (*e.g.*, CLIP)-guided masking and targets. However, they ignore the potential role of the self-training (student) model in giving feedback to the teacher for masking and targets. In this work, we present to integrate Collaborative Masking and Targets for boosting Masked AutoEncoders, namely CMT-MAE. Specifically, CMT-MAE leverages a simple collaborative masking mechanism through linear aggregation across attentions from both teacher and student models. We further propose using the output features from those two models as the collaborative target of the decoder. Our simple and effective framework pre-trained on ImageNet-1K achieves state-of-the-art linear probing and fine-tuning performance. In particular, using ViT-base, we improve the fine-tuning results of the vanilla MAE from 83.6% to 85.7%.

Introduction

Masked autoencoders (MAE) (He et al. 2021) have recently achieved advanced success in learning meaningful visual representations for many downstream tasks, *e.g.*, image classification, object detection, and semantic segmentation. Meanwhile, researchers also introduced diverse masking pipelines to show the effectiveness of masked modeling in learning meaningful representations from video (Tong et al. 2022; Feichtenhofer et al. 2022), audio (Huang et al. 2022), and MRI/CT scans (Chen et al. 2023).

Exploring masking strategies (*i.e.*, pretext tasks) and supervision targets is critical for MAE (He et al. 2021) to capture meaningful features during pre-training. In this work, we aim to simultaneously improve the powerfulness of the pre-text task and target in MAE-based pre-training on images for self-supervised vision representation learning, which boosts the performance of several downstream tasks compared to MAE (He et al. 2021) and DINO (Caron et al. 2021), as shown in Figure 1.

Early works (Bao, Dong, and Wei 2021; Atito, Awais, and Kittler 2021; He et al. 2021) on masked image modeling (MIM) mainly applied custom-designed (*i.e.*, random, block-wise, square-wise) masking during pre-training. For instance, BEiT (Bao, Dong, and Wei 2021) used a block-wise masking strategy to reconstruct discrete tokens of masked image patches for pre-training transferrable visual representations. To simplify masked image encoding, the seminal work, MAE (He et al. 2021) directly reconstructed missing pixels of 75% masked patches. MaskFeat (Wei et al. 2022a) studied five different types of features as supervision targets and observed that Histograms of Oriented Gradients (HOG) with local contrast normalization achieved the best performance. SimMIM (Xie et al. 2022) applied a large masked patch size to randomly mask the input image, where RGB values of raw pixels are recovered by a one-layer prediction head.

Recent researchers (Li et al. 2021b; Shi et al. 2022; Wei et al. 2022b; Li et al. 2022) started to leverage a teacher network or adversarial learning to generate the mask and su-

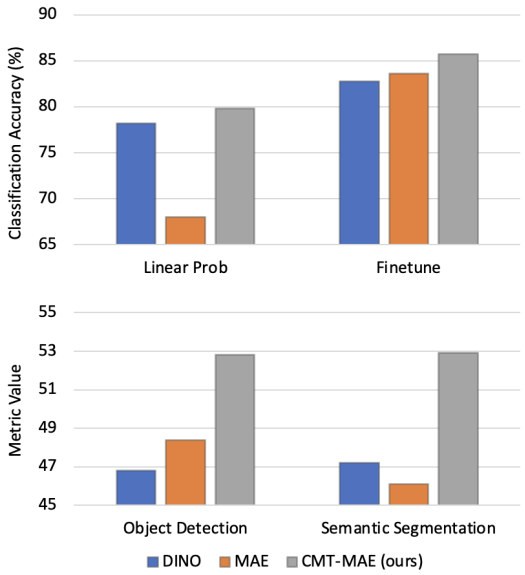

Figure 1: Comparison of our CMT-MAE with MAE and DINO on pre-trained ViT-B/16. Our method significantly outperforms previous baselines in terms of all downstream tasks.

pervision target. Benefiting from CLIP (Radford et al. 2021) pre-trained on 400 million image-language pairs, MVP (Wei et al. 2022b) introduced knowledge from CLIP as guidance to achieve impressive gains for MIM-based self-supervised visual pre-training. AttMask (Kakogeorgiou et al. 2022) applied an attention map generated from a teacher transformer encoder, *i.e*., iBoT (Zhou et al. 2022) to guide masking for the student pre-training. Similarly, SemMAE (Li et al. 2022) started to integrate semantic-guided masks from a self-supervised part-learning module with diversity constraints on attention.

While the aforementioned methods achieve promising results, they ignore the potential role of the self-training (student) model in cooperating with the teacher for collaborative masking and targets. The main challenge is that the teacher and students have different knowledge levels. The teacher network (*e.g*., CLIP) masters the knowledge at a higher level than students at initial training, while the student network starts to increase its level across the training. To address the aforementioned challenge, our key idea is to simultaneously incorporate two different knowledge-level models (*i.e*., teacher and student) to guide masking and generate reconstruction targets. During training, we aim to leverage students with self-training knowledge to help the teacher with fixed knowledge to guide MIM-based image pre-training dynamically and powerfully.

To this end, we propose a novel masked autoencoder that can integrate the student and teacher networks for collaborative masking and targets, namely CMT-MAE. In particular, we introduce a simple collaborative masking mechanism through linear aggregation across attention maps from both teacher and student models, which improves the powerfulness of guided masks. To further boost the performance of downstream tasks, the proposed framework selects representations generated from those two models as the collaborative target for the decoder during pre-training.

Our pre-training process is composed of two stages: In the first stage, a teacher transformer encoder (*i.e*., CLIP) takes an input image to extract an attention map from the last attention layer to guide masking. The student encoder generates features from unmasked patches, which are concatenated with masked tokens to feed into a decoder for recovering the teacher features of masked patches. In the second stage, we apply a student momentum encoder to generate a student-guided attention map and linearly aggregate it with a teacher-guided attention map to produce the collaborative attention map with a collaborative ratio for collaborative masking. Then masked tokens concatenate with features of unmasked patches from the student encoder to feed into the decoder. Finally, two predicted heads are linearly applied to reconstruct the teacher and student features of masked patches for collaborative targets. It should be noted that the collaborative ratio is also applied to calculate collaborative losses from the teacher and student targets.

Experimental results on ImageNet-1K, MS-COCO, ADE-20K, and DAVIS 2017 demonstrate the state-of-the-art performance of our CMT-MAE. In particular, using the backbone of the ViT-base, we improve the fine-tuning results of the vanilla MAE from 83.6% to 85.7%, and linear probing from 68.0% to 79.8%. Our method also achieves +4.8 mIoU (*i.e.,* 48.1 $\rightarrow$ 52.9) on ADE20K semantic segmentation, +6.6 $(\mathcal{J}\&\mathcal{F})_m$ (*i.e.,* 51.0 $\rightarrow$ 57.6) on DAVIS video segmentation, +2.5 AP^{box} (*i.e.,* 50.3 $\rightarrow$ 52.8) on COCO object detection, and +0.8 AP^{mask} (*i.e.,* 44.9 $\rightarrow$ 45.7) on COCO instance segmentation. In addition, qualitative visualizations of collaborative attention vividly showcase the effectiveness of our CMT-MAE in learning meaningful representations. Extensive ablation studies also demonstrate the importance of collaborative masking and collaborative targets in learning masked autoencoders for improving downstream performance.

Our main contributions can be summarized as follows:

- We present a simple yet effective masked autoencoder that can achieve collaborative masking and targets, called CMT-MAE, for boosting MIM-based visual pre-training.
- We propose a novel collaborative masking mechanism through linear aggregation across attention maps from both teacher and student networks to achieve powerful guidance.
- Extensive experiments comprehensively demonstrate the state-of-the-art superiority of our CMT-MAE over previous baselines on downstream tasks.

Related Work

Self-supervised Visual Learning. Self-supervised visual learning aims to mine the internal characteristics from images without annotations by applying well-designed pretext tasks. Early non-transformer researchers introduced instance-level (Wu et al. 2018; Chen et al. 2020a,b; Grill et al. 2020; He et al. 2020; Chen et al. 2020c; Chen and He 2021; Zbontar et al. 2021; Wu et al. 2023a; Mo, Sun, and Li 2023c,a) and cluster-based (Caron et al. 2020; Li et al. 2021a; Wang, Liu, and Yu 2021; Mo, Sun, and Li 2021, 2022) contrastive learning to pull representations from positive samples closer while pushing away features from negative pairs. Recently, contrastive learning has been widely used in self-supervised vision transformers (Chen, Xie, and He 2021; Xie et al. 2021; Caron et al. 2021; Mo, Sun, and Li 2023b; Mo and Yun 2024; Mo and Tong 2024) to achieve promising performance on visual downstream tasks. Typically, MoCov3 (Chen, Xie, and He 2021) introduced a momentum encoder in ViT (Dosovitskiy et al. 2021) to minimize the distance between representations of two augmented views from the base encoder and momentum one. To capture the local-to-global alignment, DINO (Caron et al. 2021) used a momentum encoder with multi-crop training to achieve knowledge distillation in the vision transformer. In this work, our main focus is to learn meaningful visual representations in self-supervised transformers through another acclaimed technique, *i.e*., masked image modeling.

Masked Image Modeling. Masked image modeling (MIM) has been explored in many previous works (Bao, Dong, and Wei 2021; Atito, Awais, and Kittler 2021; He et al. 2021; Wei et al. 2022a; Xie et al. 2022; Wu and Mo 2022; Wu et al. 2023b, 2024) to reconstruct the masked image patch given the unmasked counterpart as clues. Early MIM approaches (Bao, Dong, and Wei 2021; Atito, Awais, and Kit-

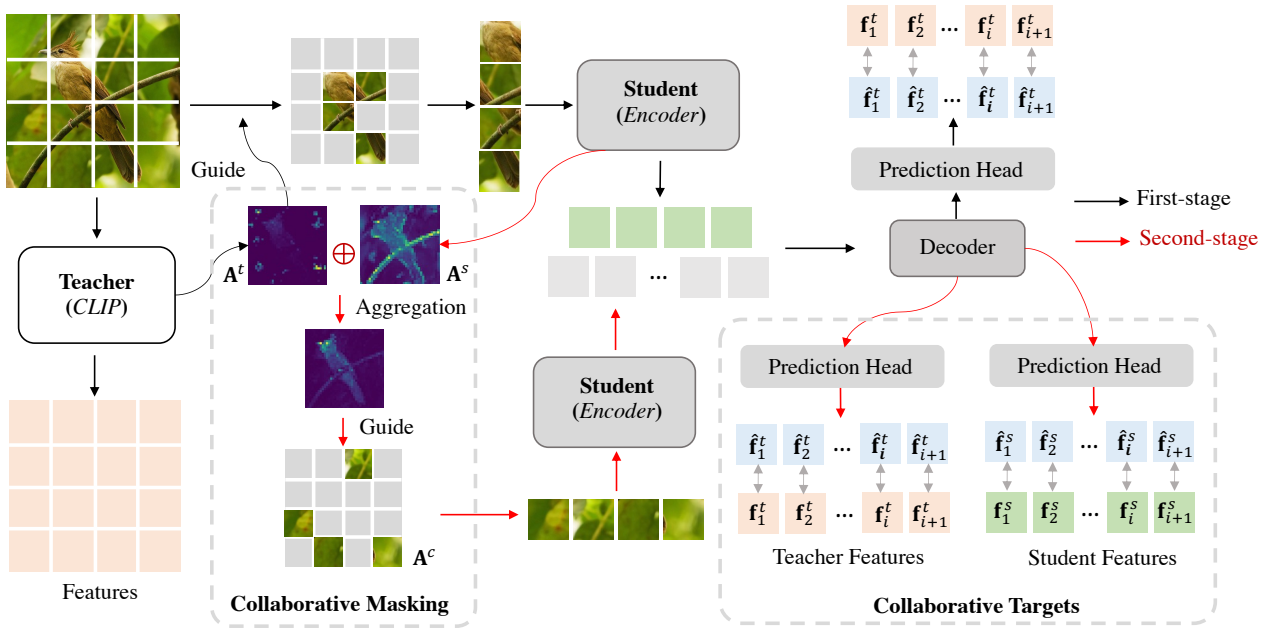

Figure 2: Illustration of the proposed Masked Autoencoder with Collaborative Masking and Targets (CMT-MAE) framework. **First-stage:** a teacher transformer encoder (*i.e.*, CLIP) takes an input image to extract an attention map $\mathbf{A^t}$ from the last attention layer to guide masking. The student encoder generates features from unmasked patches, which are concatenated with masked tokens to feed into a decoder for recovering the teacher features $\mathbf{f}_i^t$ of masked patches. **Second-stage:** a student momentum encoder takes the input image to generate a student-guided attention map $\mathbf{A^s}$, and linearly aggregates with a teacher-guided attention map $\mathbf{A^t}$ to produce the collaborative attention map $\mathbf{A^c}$ with a collaborative ratio α for collaborative masking. Then masked tokens concatenate with features of unmasked patches from the student encoder to feed into the decoder. Finally, two predicted heads are linearly applied to reconstruct the teacher features $\mathbf{f}_i^t$ and student features $\mathbf{f}_i^s$ of masked patches for collaborative targets. Note that the collaborative ratio α is also applied to calculate collaborative losses from the teacher and student targets.

tler 2021; He et al. 2021; Li et al. 2021b; Shi et al. 2022) designed customized masking strategies (*e.g.*, random, block-wise) as pre-text tasks during pre-training. For example, block-wise masking was introduced in BEiT (Bao, Dong, and Wei 2021) to learn transferrable visual representations by recovering discrete tokens of masked image patches. Given features extracted from the 25% unmasked patches, the seminal work, MAE (He et al. 2021) directly reconstructed missing pixels of 75% masked patches.

Due to the benefit of open-sourced large-scale models, MVP (Wei et al. 2022b) adopted CLIP (Radford et al. 2021) pre-trained on 400 million image-language pairs to obtain multi-modal knowledge from images as guidance for supervision. Based on iBoT (Zhou et al. 2022), a distillation-based model, AttMask (Kakogeorgiou et al. 2022) extracted an attention map from the teacher transformer encoder to guide mask generation for pre-training, while SemMAE (Li et al. 2022) integrated semantic-guided masks from a self-supervised part-learning module with diversity constraints on attention. However, they do not involve the self-training encoder itself (*i.e.*, student) with dynamic knowledge from training data to help the teacher (*e.g.*, CLIP) with fixed knowledge to guide mask and target generation for visual pre-training. In contrast, we develop a novel collaborative masking mechanism to achieve dynamic and powerful guidance through a simple yet effective linear aggregation across attention maps from both teacher and student networks, which is not addressed before yet.

Method

Given an image with masked and unmasked patches, our target is to train a masked autoencoder framework with an encoder and a decoder to recover the masked patches using unmasked counterparts. We present a simple yet effective masked autoencoder with collaborative masking and targets, named CMT-MAE, which mainly consists of two modules, Collaborative Masking and Collaborative Targets.

Preliminaries

In this section, we first describe the problem setup and notations, and then revisit the masked image modeling in MAE (He et al. 2021) and teacher-guided MAE for self-supervised visual pre-training.

Problem Setup and Notations. Given an image with a dimension of $3 \times H \times W$ and a patch resolution of P, our goal is to learn a masked autoencoder framework with an encoder $f_e(\cdot)$ and a decoder $f_d(\cdot)$ to recover the masked patches using unmasked ones. We formally denote patch

embeddings of raw input via each linear projection layer, *i.e.*, $\mathbf{x} \in \mathbb{R}^{N \times D}$, H and W are the height and width of each image, and D is the dimension of features. Note that $N = H/P \times W/P$ and N is the total number of patches.

Revisit Masked Autoencoder. To address the masked image modeling problem, MAE (He et al. 2021) first applied a random masking set M along the total number of patches, and then an encoder to extract features from unmasked patches. Finally, unmasked embeddings and masked tokens were concatenated into a decoder to recover the raw pixels of masked patches. The vanilla masking loss for each image is calculated with the mean square loss between the targeted $\mathbf{p}_i$ and predicted normalized pixels $\hat{\mathbf{p}}_i$ as:

$$\mathcal{L}_{\text{mae}} = \frac{1}{|M|} \sum_{i \in M} ||\mathbf{p}_i - \hat{\mathbf{p}}_i||_2^2 \tag{1}$$

where $|M|$ denotes the total number of masked patches in the masking set M.

Revisit Teacher-guided Masked Autoencoder. Based on the aforementioned seminal work, recent researchers (Wei et al. 2022b; Kakogeorgiou et al. 2022; Li et al. 2022) started to use off-the-shelf large-scale pre-trained models (*e.g.*, CLIP (Radford et al. 2021)) as a teacher to guide mask generation and supervision targets. Specifically, they applied an attention map $\mathbf{A^t}$ extracted from a teacher transformer encoder to generate the masking set M^t, and representations as the reconstruction target. The teacher-guided masking loss for each image is computed with the mean square loss between the targeted $\mathbf{f}_i^t$ and predicted normalized features $\hat{\mathbf{f}}_i^t$ as:

$$\mathcal{L}_{\text{teacher-mae}} = \frac{1}{|M^t|} \sum_{i \in M^t} ||\mathbf{f}_i^t - \hat{\mathbf{f}}_i^t||_2^2 \tag{2}$$

where $|M^t|$ denotes the total number of masked patches in the teacher-guided masking set M^t.

However, such a training objective will pose the main challenge for collaborative masking and targets. These approaches do not explicitly incorporate the self-training (*i.e.*, student) model to cooperate with the teacher during pre-training. Furthermore, the teacher and students have different knowledge levels: the teacher network (*e.g.*, CLIP) has a higher-level knowledge than students at initial training; the student network starts to aggrandize its knowledge with more training epochs. To address the challenge, we propose a simple yet effective masked autoencoder that can achieve collaborative masking and targets, called CMT-MAE, as illustrated in Figure 2.

Collaborative Masking

In order to explicitly achieve collaborative masking guided by both the teacher and student, we propose a two-stage training paradigm with a simple collaborative masking mechanism through linear aggregation across attention maps from both models to improve the powerfulness of generated masks. Specifically, in the first stage, we leverage a teacher transformer encoder (*i.e.*, CLIP (Radford et al. 2021)) takes an input image to extract an attention map $\mathbf{A^t}$ from the last attention layer to guide masking. The student encoder generates features from unmasked patches, which are concatenated with masked tokens to feed into a decoder for recovering the teacher features $\mathbf{f}_i^t$ of masked patches.

Since momentum encoders (Tarvainen and Valpola 2017; He et al. 2020) is a technique often used in self-supervised and semi-supervised learning to obtain slow-moving target representations, leading to more stable self-training and enhanced representations. In the second stage, as shown in Figure 2, we leverage a student momentum encoder to take the input image to generate a student-guided attention map $\mathbf{A^s}$. Following (Tarvainen and Valpola 2017; **?**), we update the student momentum encoder using an exponential moving average of the corresponding online encoders with coefficient m.[1] The student $\mathbf{A^s}$ and teacher $\mathbf{A^t}$ attention maps are then linearly aggregated into a final collaborative map $\mathbf{A^c}$ of the form as

$$\mathbf{A^c} = \alpha * \mathbf{A^s} + (\mathbf{1} - \alpha) * \mathbf{A^t} \tag{3}$$

where α denotes the collaborative ratio. The collaborative map $\mathbf{A^c}$ is followingly applied to generate a masking set M^c to split the input image into masked and unmasked patches for the second training stage.

Collaborative Targets

Beyond collaborative masking, we introduce collaborative targets as the training objective in the second stage to dynamically select representations from both the teacher and student for simultaneous optimization. With a masking set M^c guided by the collaborative attention map $\mathbf{A^c}$, we concatenate masked tokens with features of unmasked patches from the student encoder to feed into the decoder. Since updates to the student encoders are slowly incorporated into the momentum encoders, the target representations display smoother behavior during the training process. In the end, we apply two predicted heads to linearly reconstruct the teacher features $\mathbf{f}_i^t$ and student features $\mathbf{f}_i^s$ of masked patches for collaborative targets.

The overall objective of our model in the second stage is simply optimized in an end-to-end manner as:

$$\mathcal{L}_{\text{cmt-mae}} = \frac{1}{|M^c|} \sum_{i \in M^c} \alpha * ||\mathbf{f}_i^s - \hat{\mathbf{f}}_i^s||_2^2 + (1-\alpha) * ||\mathbf{f}_i^t - \hat{\mathbf{f}}_i^t||_2^2 \tag{4}$$

where $|M^c|$ denotes the total number of masked patches in the collaborative masking set M^c. $\mathbf{f}_i^s, \hat{\mathbf{f}}_i^s$ denote the target and prediction of student features, and $\mathbf{f}_i^t, \hat{\mathbf{f}}_i^t$ for teacher features. α is the collaborative ratio. Note that the collaborative ratio α is applied to calculate collaborative losses from the teacher and student targets.

Experiments

Experimental setup

Datasets. Following previous methods (He et al. 2021), we use ImageNet-1K (Deng et al. 2009) for image classification, MS-COCO (Lin et al. 2014) for object detection

[1] EMA update is $\hat{\theta} \leftarrow \hat{\theta} + (1-m)\theta$, where θ and $\hat{\theta}$ are the parameters of online and momentum encoders.

Method	Pre-train Dataset	Pre-train Epochs	ViT-B/16		ViT-L/16	
			Linear Probing	Fine-tuning	Linear Probing	Fine-tuning
Training from scratch						
DeiT (Touvron et al. 2020)	–	–	–	81.8	–	–
ViT (Dosovitskiy et al. 2021)	–	–	–	82.3	–	84.5
Contrastive-based Pre-Training						
AttMask (Kakogeorgiou et al. 2022)	ImageNet-1K	100	75.7	–	–	–
DINO (Caron et al. 2021)	ImageNet-1K	300	78.2	82.8	–	–
MoCo v3 (Chen, Xie, and He 2021)	ImageNet-1K	300	76.5	83.2	–	84.1
iBOT (Zhou et al. 2022)	ImageNet-1K	1600	79.5	84.0	81.0	84.8
MIM-based Pre-Training						
BEiT (Bao, Dong, and Wei 2021)	ImageNet-1K	800	56.7	83.2	–	–
MAE (He et al. 2021)	ImageNet-1K	1600	68.0	83.6	75.1	85.9
MaskFeat (Wei et al. 2022a)	ImageNet-1K	1600	68.0	84.0	–	85.7
SimMIM (Xie et al. 2022)	ImageNet-1K	800	56.7	83.8	–	–
PeCo (Dong et al. 2021)	ImageNet-1K	300	–	84.1	–	–
MVP (Wei et al. 2022b)	ImageNet-1K	300	–	84.4	–	86.3
data2vec (Baevski et al. 2022)	ImageNet-1K	1600	–	84.2	–	86.6
SemMAE (Li et al. 2022)	ImageNet-1K	800	68.7	84.5	–	–
CMT-MAE (ours)	ImageNet-1K	800	**79.8**	**85.7**	**81.6**	**87.2**

Table 1: ImageNet-1K image classification. We performed a linear probing and fine-tuning on pre-trained ViT-B/16 and ViT-L/16 models for image classification on ImageNet-1K benchmark. We report the top-1 accuracy to evaluate the quality of pre-trained representations. The best results are indicated in **bold** numbers.

and instance segmentation, and ADE20K (Zhou et al. 2017, 2018) for semantic segmentation. We closely follow previous work (Chen, Xie, and He 2021; Xie et al. 2021; Caron et al. 2021), and adopt the Mask R-CNN (He et al. 2017) as the detector. The ViT-Base (Dosovitskiy et al. 2021) backbone weights are initialized with weights pre-trained on ImageNet-1K using our CMT-MAE. Other settings are the same as the implementation in this work (He et al. 2021). Following the settings in (He et al. 2021; Bao, Dong, and Wei 2021), we use the UPerNet approach (Xiao et al. 2018) based on our ImageNet-1K pre-trained ViT-Base for evaluation. For a fair comparison, we fine-tune the detector with the same learning rate in (He et al. 2021; Bao, Dong, and Wei 2021). For video object segmentation, we use DAVIS-2017 dataset (Pont-Tuset et al. 2017) that includes 60 training, 30 validation, and 60 testing videos.

Evaluation Metrics. We follow previous masked image modeling work (He et al. 2021; Bao, Dong, and Wei 2021) to report the classification accuracy of linear probing and fine-tuning. For object detection and instance segmentation on MS-COCO, we apply AP^b and AP^m as metrics for the bounding boxes and the instance masks. mIoU results are reported to evaluate semantic segmentation on ADE20K. For video object segmentation on DAVIS-2017, we use Jabri-based $(\mathcal{J}\&\mathcal{F})_m$, $\mathcal{J}_m$, $\mathcal{F}_m$ as metrics to evaluate segmenting frames based on the nearest neighbor between consecutive scenes.

Implementation. For input images, the resolution is resized to 224×224, *i.e.*, $H = W = 224$. We follow prior work (He et al. 2021) and apply a patch size of 16, *i.e.*, $P = 16$. The base and large models of ViT (Dosovitskiy et al. 2021) architecture are used for experiments. We mask 75% on patches of each image, same as in MAE (He et al. 2021). The model is pre-trained on ImageNet-1K for 800 epochs with the AdamW (Loshchilov and Hutter 2019) optimizer with a learning rate of 1.5e-4, a decay rate of 0.05, and a batch size of 4096. For fine-tuning on ImageNet-1K, the model is trained for 100 epochs with a batch size of 256. For the MS-COCO dataset, we train the model for 12 epochs with a batch size of 16, and an initial learning rate of 2e-4. The learning rate is decayed by 10 at epochs 8 and 11. For the ADE20K dataset, the model is trained for 160K iterations with an initial learning rate of 3e-5 and a layer-wise learning rate decay of 0.9. We set the weight decay to 0.05 and the drop path rate to 0.1. Multi-scale testing is not used for a fair comparison with precious approaches (He et al. 2021; Bao, Dong, and Wei 2021).

Comparison to prior work

In this work, we propose a novel and effective framework with MAE pre-training for downstream tasks, *i.e.*, linear probing, fine-tuning, object detection, instance segmentation, semantic segmentation, and video object segmentation. To validate the effectiveness of the proposed CMT-MAE, we comprehensively compare it to previous baselines, including training from scratch (Touvron et al. 2020; Dosovitskiy et al. 2021), contrastive-based pre-training (Kakogeorgiou et al. 2022; Caron et al. 2021; Chen, Xie, and He 2021; Zhou et al. 2022), and previous MIM-based pre-training (Bao, Dong, and Wei 2021; He et al. 2021; Wei et al. 2022a; Xie et al. 2022; Dong et al. 2021; Wei et al. 2022b; Baevski et al. 2022; Li et al. 2022) approaches.

Image classification. Table 1 reports the quantitative comparison results of linear probing and fine-tuning on pre-

Method	AP^{box}	AP^{mask}	mIoU
Supervised Training			
DeiT (Touvron et al. 2020)	47.9	42.9	47.4
Contrastive-based Pre-Training			
DINO (Caron et al. 2021)	46.8	41.5	47.2
MoCo v3 (Chen, Xie, and He 2021)	47.9	42.7	47.3
MIM-based Pre-Training			
BEiT (Bao, Dong, and Wei 2021)	42.1	37.8	45.8
MAE (He et al. 2021) (800epoch)	48.4	42.6	46.1
PeCo (Dong et al. 2021)	43.9	39.8	46.7
SplitMask (El-Nouby et al. 2021)	46.8	42.1	45.7
iBoT (Zhou et al. 2022)	48.2	42.7	50.0
CAE (Chen et al. 2022)	49.2	43.3	48.8
MVP (Wei et al. 2022b)	–	–	52.4
SemMAE (Li et al. 2022)	–	–	46.3
MAE (He et al. 2021) (1600epoch)	50.3	44.9	48.1
CMT-MAE (ours)	**52.8**	**45.7**	**52.9**

Table 2: COCO object detection, instance segmentation, and ADE20K semantic segmentation. We fine-tuned pre-trained ViT-B/16 models to perform COCO object detection, instance segmentation, and ADE20K semantic segmentation. The AP^{box}, AP^{mask}, and mIoU metrics denote the results of COCO detection, COCO segmentation, and ADE20K segmentation, respectively. The best results are indicated in **bold** numbers.

trained ViT-B/16 and ViT-L/16 models. As can be seen, we achieve the best performance in terms of all metrics for both models. In particular, the proposed CMT-MAE significantly outperforms MAE (He et al. 2021), the original masked image modeling baseline, by 11.8% & 2.1% and 6.5% & 1.3% relative top-1 accuracies in terms of linear probing & fine-tuning on ViT-B/16 and ViT-L/16 models. Moreover, we achieve superior performance gains compared to SemMAE (Li et al. 2022), the recent MIM-based pre-training approach that applied semantic-guided masks and diversity constraints on attention. Meanwhile, our CMT-MAE outperforms iBoT (Zhou et al. 2022) by 1.7% and 2.4% relative top-1 accuracies in terms of fine-tuning on ViT-B/16 and ViT-L/16 models. The proposed CMT-MAE also achieves better results than DINO (Caron et al. 2021), a strong contrastive-based pre-training baseline. These significant improvements demonstrate the superiority of our method in learning better representations during pre-training for image classification.

Object detection & instance segmentation. For the COCO object detection & instance segmentation benchmarks, we report the quantitative comparison results of COCO object detection and instance segmentation on the pre-trained ViT-B/16 model in Table 2. We can observe that the proposed CMT-MAE achieves the best results in terms of all metrics. Compared to MAE (He et al. 2021) trained on 1600 epochs, we achieve performance gains of 2.5@AP^{box} and 0.8@AP^{mask}. We also achieve highly better results than other contrastive-based (Caron et al. 2021) and MIM-based (Bao, Dong, and Wei 2021; Li et al. 2022) pre-training approaches.

Method	Backbone	$(\mathcal{J}\&\mathcal{F})_m$	$\mathcal{J}_m$	$\mathcal{F}_m$
MAE (He et al. 2021)	ViT-B/16	51.0	49.4	52.6
I-JEPA (Assran et al. 2023)	ViT-B/16	56.2	56.1	56.3
CMT-MAE (ours)	ViT-B/16	**57.6**	**56.7**	**58.5**
MAE (He et al. 2021)	ViT-L/16	53.4	52.5	54.3
I-JEPA (Assran et al. 2023)	ViT-L/16	56.6	56.3	56.9
CMT-MAE (ours)	ViT-L/16	**60.5**	**59.7**	**61.3**

Table 3: DAVIS video object segmentation. We performed DAVIS 2017 video object segmentation using ViT-B/16 and ViT-L/16 pre-trained on ImageNet-1K. We report Jabri-based metrics $(\mathcal{J}\&\mathcal{F})_m$, $\mathcal{J}_m$, $\mathcal{F}_m$ to evaluate the quality of frozen pre-trained representations. The best results are indicated in **bold** numbers.

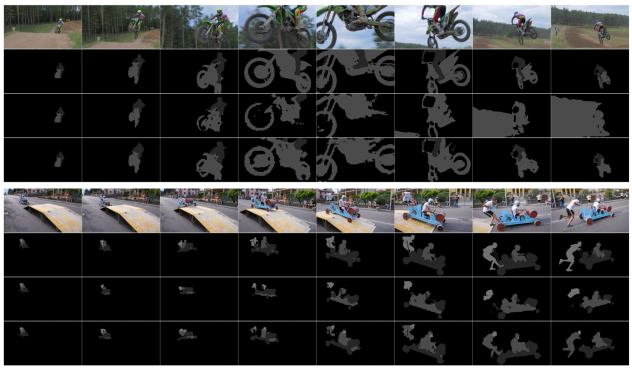

Figure 3: Visualizations of DAVIS 2017 video object segmentation. Four rows for each case represent raw frames, ground-truth masks, MAE predictions, and our CMT-MAE predictions. We visualize the segmentation masks of DAVIS 2017 video object segmentation using ViT-B/16 pre-trained on ImageNet-1K. The proposed CMT-MAE produces much more accurate and high-quality segmentation masks.

Semantic segmentation. Table 2 also shows the quantitative comparison results of ADE20K semantic segmentation on the ViT-B/16 model pre-trained on ImageNet-1K. Our CMT-MAE significantly outperforms MAE (He et al. 2021) by 4.8@mIoU and also achieves better performance than MVP (Wei et al. 2022b), the strong baseline using CLIP (Radford et al. 2021) knowledge pre-trained on 400 million image-text pairs. These results further validate the effectiveness of our collaborative masking and collaborative masking in learning discriminative embeddings across pre-training for semantic segmentation.

Video object segmentation. We also present additional video object segmentation on the DAVIS 2017 benchmark using ImageNet-1K pre-trained ViT-B/16 and ViT-L/16 models, as shown in Table 3. We achieve superior performance gains of 6.6@$(\mathcal{J}\&\mathcal{F})_m$, 7.3@$\mathcal{J}_m$, 5.9@$\mathcal{F}_m$, and 7.1@$(\mathcal{J}\&\mathcal{F})_m$, 7.2@$\mathcal{J}_m$, 7.0@$\mathcal{F}_m$ in terms of pre-trained ViT-B/16 and ViT-L/16, compared to MAE (He et al. 2021). We also achieve better results than I-JEPA (Assran et al. 2023), the recent joint embedding predictive architecture for image self-supervised learning. These qualitative results

CM	CT	Linear Probing	Fine-tuning	AP^{box}	AP^{mask}	mIoU	$(\mathcal{J}\&\mathcal{F})_m$	$\mathcal{J}_m$	$\mathcal{F}_m$
✗	✗	68.0	83.6	48.4	42.6	46.1	51.0	49.4	52.6
✓	✗	73.5	84.2	50.3	43.5	48.3	53.8	51.8	55.8
✗	✓	74.2	84.5	50.9	44.2	49.2	54.6	52.9	56.3
✓	✓	**79.8**	**85.7**	**52.8**	**45.7**	**52.9**	**57.6**	**56.7**	**58.5**

Table 4: Ablation studies on component analysis. We performed ablation studies on Collaborative Masking (CM) and Collaborative Targets (CT) modules using a pre-trained ViT-B/16 on ImageNet-1K. The best results are indicated in **bold** numbers.

α	Linear Probing	Fine-tuning	AP^{box}	AP^{mask}	mIoU	$(\mathcal{J}\&\mathcal{F})_m$	$\mathcal{J}_m$	$\mathcal{F}_m$
0%	77.1	84.6	51.7	44.7	51.5	55.3	54.0	56.6
10%	78.7	85.1	52.1	45.2	52.1	56.5	55.6	57.4
30%	**79.8**	**85.7**	**52.8**	**45.7**	**52.9**	**57.6**	**56.7**	**58.5**
50%	79.3	85.5	52.5	45.5	52.5	57.2	56.1	58.3
70%	78.9	85.2	52.2	45.3	52.3	56.8	55.8	57.8
90%	78.3	84.9	51.9	45.1	51.9	56.2	55.3	57.1
100%	77.6	84.7	51.8	44.9	51.7	55.8	54.7	56.9

Table 5: Ablation studies on collaborative ratio α. We performed ablation studies using an ImageNet-1K pre-trained ViT-B/16 model to explore impacts on collaborative ratio. The best results are indicated in **bold** numbers.

also showcase the effectiveness of our CMT-MAE in generating much more accurate and high-quality segmentation masks on video objects compared to MAE (He et al. 2021), as shown in Figure 3.

Experimental analysis

In this section, we performed ablation studies to demonstrate the benefit of introducing Collaborative Masking (CM) and Collaborative Targets (CT) modules. We also conducted extensive experiments to explore the impact of collaboration ratio α and learned meaningful collaborative attention maps.

Collaborative Masking & Collaborative Targets. In order to validate the effectiveness of the introduced collaborative masking (CM) and collaborative targets (CT), we ablate the necessity of each module and report the quantitative results in Table 4. We can observe that adding CM to the vanilla baseline highly increases the results of 5.5@Linear Probing, 0.6@Fine-tuning, 1.9@AP^{box}, 0.9@AP^{mask}, 2.2@mIoU, 2.8@$(\mathcal{J}\&\mathcal{F})_m$, 2.4@$\mathcal{J}_m$, 3.2@$\mathcal{F}_m$, which demonstrates the benefit of collaborative masking in generating meaningful representations guided by both the teacher and student during pre-training. Meanwhile, introducing only CT in the baseline also increases the downstream performance in terms of all metrics. More importantly, incorporating CM and CT together into the baseline significantly raises the performance by 11.8 @Linear Probing, 2.1@Fine-tuning, 4.4@AP^{box}, 3.1@AP^{mask}, 6.8@mIoU, 6.6@$(\mathcal{J}\&\mathcal{F})_m$, 7.3@$\mathcal{J}_m$, 5.9@$\mathcal{F}_m$. These improving results validate the importance of collaborative masking and targets in learning collaborative representations from both teacher and student for masked autoencoders.

Trade-off on Collaborative Ratio. The number of collaborative ratios in the proposed collaborative masking and targets affects the pre-trained representations for diverse downstream tasks. To explore such effects more comprehensively, we varied the number of ratios from $\{0\%, 10\%, 30\%, 50\%, 70\%, 90\%, 100\%\}$. The comparison results of all downstream tasks using a ViT-B/16 model pre-trained on ImageNet-1K are reported in Table 5. When the number of collaboration ratio α is 30%, we achieve the best downstream performance in terms of all metrics. With the increase of collaboration ratio from 0% to 30%, the proposed CMT-MAE consistently raises results, which shows the importance of collaborative masking and collaborative targets in masked autoencoders for learning discriminative representations. However, increasing the collaboration ratio from 30% to 90% will not continually improve the result since there might be a trade-off between the teacher and student to learn different representations during pre-training.

Conclusion

In this work, we present CMT-MAE, a simple yet effective masked autoencoder that can simultaneously achieve collaborative masking and targets. We leverage a novel collaborative masking mechanism through linear aggregation across attentions from both teacher and student models. We further use their representations as the collaborative target of the decoder for reconstruction. Experimental results on ImageNet-1K, MS-COCO, ADE-20K, and DAVIS 2017 validate the state-of-the-art superiority of the proposed framework. In addition, qualitative visualizations vividly showcase the effectiveness of our CMT-MAE in capturing meaningful representations for downstream tasks. Extensive ablation studies also demonstrate the importance of collaborative masking and collaborative targets in learning masked autoencoders for improving downstream performance.

References

Assran, M.; Duval, Q.; Misra, I.; Bojanowski, P.; Vincent, P.; Rabbat, M.; LeCun, Y.; and Ballas, N. 2023. Self-Supervised Learning from Images with a Joint-Embedding Predictive Architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 15619–15629.

Atito, S.; Awais, M.; and Kittler, J. 2021. SiT: Self-supervised vIsion Transformer. *arXiv preprint arXiv:2104.03602*.

Baevski, A.; Hsu, W.; Xu, Q.; Babu, A.; Gu, J.; and Auli, M. 2022. data2vec: A General Framework for Self-supervised Learning in Speech, Vision and Language. In Chaudhuri, K.; Jegelka, S.; Song, L.; Szepesvári, C.; Niu, G.; and Sabato, S., eds., *Proceedings of International Conference on Machine Learning (ICML)*, volume 162 of *Proceedings of Machine Learning Research*, 1298–1312. PMLR.

Bao, H.; Dong, L.; and Wei, F. 2021. BEiT: BERT Pre-Training of Image Transformers. *arXiv preprint arXiv:2106.08254*.

Caron, M.; Misra, I.; Mairal, J.; Goyal, P.; Bojanowski, P.; and Joulin, A. 2020. Unsupervised Learning of Visual Features by Contrasting Cluster Assignments. In *Proceedings of Advances in Neural Information Processing Systems*.

Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; and Joulin, A. 2021. Emerging Properties in Self-Supervised Vision Transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020a. A Simple Framework for Contrastive Learning of Visual Representations. In *Proceedings of International Conference on Machine Learning (ICML)*.

Chen, T.; Kornblith, S.; Swersky, K.; Norouzi, M.; and Hinton, G. 2020b. Big Self-Supervised Models are Strong Semi-Supervised Learners. In *Proceedings of Advances in Neural Information Processing Systems*.

Chen, X.; Ding, M.; Wang, X.; Xin, Y.; Mo, S.; Wang, Y.; Han, S.; Luo, P.; Zeng, G.; and Wang, J. 2022. Context Autoencoder for Self-Supervised Representation Learning. *arXiv preprint arXiv:2202.03026*.

Chen, X.; Fan, H.; Girshick, R.; and He, K. 2020c. Improved Baselines with Momentum Contrastive Learning. *arXiv preprint arXiv:2003.04297*.

Chen, X.; and He, K. 2021. Exploring Simple Siamese Representation Learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Chen, X.; Xie, S.; and He, K. 2021. An Empirical Study of Training Self-Supervised Vision Transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*.

Chen, Z.; Agarwal, D.; Aggarwal, K.; Safta, W.; Balan, M. M.; Sethuraman, V. S.; and Brown, K. 2023. Masked Image Modeling Advances 3D Medical Image Analysis. In *Proceedings of IEEE Winter Conference on Applications of Computer Vision (WACV)*.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 248–255.

Dong, X.; Bao, J.; Zhang, T.; Chen, D.; Zhang, W.; Yuan, L.; Chen, D.; Wen, F.; and Yu, N. 2021. PeCo: Perceptual Codebook for BERT Pre-training of Vision Transformers. *arXiv preprint arXiv:2111.12710*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *Proceedings of International Conference on Learning Representations*.

El-Nouby, A.; Izacard, G.; Touvron, H.; Laptev, I.; Jégou, H.; and Grave, E. 2021. Are Large-scale Datasets Necessary for Self-Supervised Pre-training? *arXiv preprint arXiv:2112.10740*.

Feichtenhofer, C.; Fan, H.; Li, Y.; and He, K. 2022. Masked Autoencoders As Spatiotemporal Learners. *arXiv:2205.09113*.

Grill, J.-B.; Strub, F.; Altché, F.; Tallec, C.; Richemond, P.; Buchatskaya, E.; Doersch, C.; Avila Pires, B.; Guo, Z.; Gheshlaghi Azar, M.; Piot, B.; kavukcuoglu, k.; Munos, R.; and Valko, M. 2020. Bootstrap Your Own Latent - A New Approach to Self-Supervised Learning. In *Proceedings of Advances in Neural Information Processing Systems*.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. B. 2021. Masked Autoencoders Are Scalable Vision Learners. *arXiv preprint arXiv:2111.06377*.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum Contrast for Unsupervised Visual Representation Learning. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9729–9738.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask R-CNN. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2980–2988.

Huang, P.-Y.; Xu, H.; Li, J. B.; Baevski, A.; Auli, M.; Galuba, W.; Metze, F.; and Feichtenhofer, C. 2022. Masked Autoencoders that Listen. *arXiv:2207.06405*.

Kakogeorgiou, I.; Gidaris, S.; Psomas, B.; Avrithis, Y.; Bursuc, A.; Karantzalos, K.; and Komodakis, N. 2022. What to Hide from Your Students: Attention-Guided Masked Image Modeling. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 300–318.

Li, G.; Zheng, H.; Liu, D.; Wang, C.; Su, B.; and Zheng, C. 2022. SemMAE: Semantic-Guided Masking for Learning Masked Autoencoders. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*.

Li, J.; Zhou, P.; Xiong, C.; and Hoi, S. 2021a. Prototypical Contrastive Learning of Unsupervised Representations. In *Proceedings of International Conference on Learning Representations (ICLR)*.

Li, Z.; Chen, Z.; Yang, F.; Li, W.; Zhu, Y.; Zhao, C.; Deng, R.; Wu, L.; Zhao, R.; Tang, M.; and Wang, J. 2021b. MST:

Masked Self-Supervised Transformer for Visual Representation. In *Proceedings of Advances in Neural Information Processing Systems*.

Lin, T.-y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; and Zitnick, C. L. 2014. Microsoft COCO: Common Objects in Context. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 740–755.

Loshchilov, I.; and Hutter, F. 2019. Decoupled Weight Decay Regularization. In *Proceedings of International Conference on Learning Representations (ICLR)*.

Mo, S.; Sun, Z.; and Li, C. 2021. Siamese Prototypical Contrastive Learning. In *BMVC*.

Mo, S.; Sun, Z.; and Li, C. 2022. Rethinking Prototypical Contrastive Learning through Alignment, Uniformity and Correlation. In *BMVC*.

Mo, S.; Sun, Z.; and Li, C. 2023a. Exploring Data Augmentations on Self-/Semi-/Fully- Supervised Pre-trained Models. *arXiv preprint arXiv:2310.18850*.

Mo, S.; Sun, Z.; and Li, C. 2023b. Multi-Level Contrastive Learning for Self-Supervised Vision Transformers. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2778–2787.

Mo, S.; Sun, Z.; and Li, C. 2023c. Representation Disentanglement in Generative Models with Contrastive Learning. *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 1531–1540.

Mo, S.; and Tong, S. 2024. Connecting Joint-Embedding Predictive Architecture with Contrastive Self-supervised Learning. *arXiv preprint arXiv: 2410.19560*.

Mo, S.; and Yun, S. 2024. DMT-JEPA: Discriminative Masked Targets for Joint-Embedding Predictive Architecture. *arXiv preprint arXiv: 2405.17995*.

Pont-Tuset, J.; Perazzi, F.; Caelles, S.; Arbeláez, P.; Sorkine-Hornung, A.; and Gool, L. V. 2017. The 2017 DAVIS Challenge on Video Object Segmentation. *arXiv preprint arXiv:1704.00675*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. *arXiv preprint arXiv:2103.00020*.

Shi, Y.; Siddharth, N.; Torr, P. H.; and Kosiorek, A. R. 2022. Adversarial Masking for Self-Supervised Learning. In *Proceedings of International Conference on Machine Learning*.

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

Tong, Z.; Song, Y.; Wang, J.; and Wang, L. 2022. VideoMAE: Masked Autoencoders are Data-Efficient Learners for Self-Supervised Video Pre-Training. In *Proceedings of Advances in Neural Information Processing Systems*.

Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; and Jégou, H. 2020. Training data-efficient image transformers & distillation through attention. *arXiv preprint arXiv:2012.12877*.

Wang, X.; Liu, Z.; and Yu, S. X. 2021. CLD: Unsupervised Feature Learning by Cross-Level Instance-Group Discrimination. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Wei, C.; Fan, H.; Xie, S.; Wu, C.-Y.; Yuille, A.; and Feichtenhofer, C. 2022a. Masked Feature Prediction for Self-Supervised Visual Pre-Training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 14668–14678.

Wei, L.; Xie, L.; Zhou, W.; Li, H.; and Tian, Q. 2022b. MVP: Multimodality-Guided Visual Pre-training. In *Proceedings of the European Conference on Computer Vision*, 337–353.

Wu, J.; and Mo, S. 2022. Object-wise Masked Autoencoders for Fast Pre-training. *arXiv preprint arXiv:2205.14338*.

Wu, J.; Mo, S.; Atito, S.; Feng, Z.; Kittler, J.; and Awais, M. 2024. DailyMAE: Towards Pretraining Masked Autoencoders in One Day. *arXiv preprint arXiv:2404.00509*.

Wu, J.; Mo, S.; Atito, S.; Kittler, J.; Feng, Z.; and Awais, M. 2023a. Beyond Accuracy: Statistical Measures and Benchmark for Evaluation of Representation from Self-Supervised Learning. *arXiv preprint arXiv:2312.01118*.

Wu, J.; Mo, S.; Awais, M.; Atito, S.; Feng, Z.; and Kittler, J. 2023b. Masked Momentum Contrastive Learning for Zero-shot Semantic Understanding. *arXiv preprint arXiv:2308.11448*.

Wu, Z.; Xiong, Y.; Yu, S. X.; and Lin, D. 2018. Unsupervised Feature Learning via Non-Parametric Instance Discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Xiao, T.; Liu, Y.; Zhou, B.; Jiang, Y.; and Sun, J. 2018. Unified Perceptual Parsing for Scene Understanding. In *Proceedings of European Conference on Computer Vision (ECCV)*, 432–448.

Xie, Z.; Lin, Y.; Yao, Z.; Zhang, Z.; Dai, Q.; Cao, Y.; and Hu, H. 2021. Self-Supervised Learning with Swin Transformers. *arXiv preprint arXiv:2105.04553*.

Xie, Z.; Zhang, Z.; Cao, Y.; Lin, Y.; Bao, J.; Yao, Z.; Dai, Q.; and Hu, H. 2022. SimMIM: A Simple Framework for Masked Image Modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9653–9663.

Zbontar, J.; Jing, L.; Misra, I.; LeCun, Y.; and Deny, S. 2021. Barlow Twins: Self-Supervised Learning via Redundancy Reduction. In *Proceedings of International Conference on Machine Learning (ICML)*.

Zhou, B.; Zhao, H.; Puig, X.; Fidler, S.; Barriuso, A.; and Torralba, A. 2017. Scene Parsing through ADE20K Dataset. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5122–5130.

Zhou, B.; Zhao, H.; Puig, X.; Fidler, S.; Barriuso, A.; and Torralba, A. 2018. Semantic Understanding of Scenes Through the ADE20K Dataset. *International Journal of Computer Vision (IJCV)*, 127: 302–321.

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