MIMIC BEFORE RECONSTRUCT: ENHANCING MASKED AUTOENCODERS WITH FEATURE MIMICKING

Anonymous authors
Paper under double-blind review

ABSTRACT

Masked Autoencoders (MAE) have been popular paradigms for large-scale vision representation pre-training. However, MAE solely reconstructs the low-level RGB signals after the decoder and lacks supervision upon high-level semantics for the encoder, thus suffering from sub-optimal learned representations and long pre-training epochs. To alleviate this, previous methods simply replace the pixel reconstruction targets of 75% masked tokens by encoded features from pre-trained image-image (DINO) or image-language (CLIP) contrastive learning. Different from those efforts, we propose to **M**imic before **R**econstruct for Masked Autoencoders, named as **MR-MAE**, which jointly learns high-level and low-level representations without interference during pre-training. For high-level semantics, MR-MAE employs a mimic loss over 25% visible tokens from the encoder to capture the pre-trained patterns encoded in CLIP and DINO. For low-level structures, we inherit the reconstruction loss in MAE to predict RGB pixel values for 75% masked tokens after the decoder. As MR-MAE applies high-level and low-level targets respectively at different partitions, the learning conflicts between them can be naturally overcome and contribute to superior visual representations for various down-stream tasks. On ImageNet-1K, the MR-MAE base pre-trained for only 200 epochs achieves 85.0% top-1 accuracy after fine-tuning, surpassing MAE base pre-trained for 1600 epochs by +1.4%. Furthermore, by appending masked convolution stages, MR-MCMAE reaches 85.8%, better than previous state-of-the-art BEiT V2 base by +0.3% with much fewer computational resources (25% vs 100% tokens fed in the encoder, and 400 vs 1600 pre-training epochs). Code and pre-trained models will be released.

1 INTRODUCTION

Masked Language Modeling (MLM) (Devlin et al., 2018; Brown et al., 2020; Radford et al., 2019) has revolutionized natural language understanding via the large-scale pre-training. Motivated by this, Masked Autoencoders (MAE) (He et al., 2022b) explore how to adopt MLM paradigm into vision representation learning with a vision transformer (Dosovitskiy et al., 2020) of asymmetric encoder-decoder architectures. MAE only encodes 25% visible image tokens and reconstructs the RGB pixels values of other 75% masked tokens. The representations learned through MAE have shown promising performances on various downstream vision tasks, which surpass the contrastive learning paradigms (Radford et al., 2021b; Caron et al., 2021; He et al., 2020).

Although MAE is rising to be the dominant approaches for vision representation learning, it still suffers from the following disadvantages compared with its MLM counterparts. Firstly, the success of MLM pre-training (Devlin et al., 2018) benefits from reconstructing the human-abstracted word tokens with rich semantics. It poses a non-trivial pre-text task that guides the transformer to learn informative representations for language understanding. Different from the high-level supervisions in language modeling, the low-level RGB signals of MAE (He et al., 2022b) is too primitive and redundant, which fail to unleash the full understanding capacity of masked autoencoding on down-stream vision tasks. Secondly, MAE (He et al., 2022b) employs an asymmetric architecture with a heavy encoder and a light decoder, where the encoder is preserved after pre-training for downstream transfer learning. However, MAE only applies the pre-training supervision upon decoder's outputs,

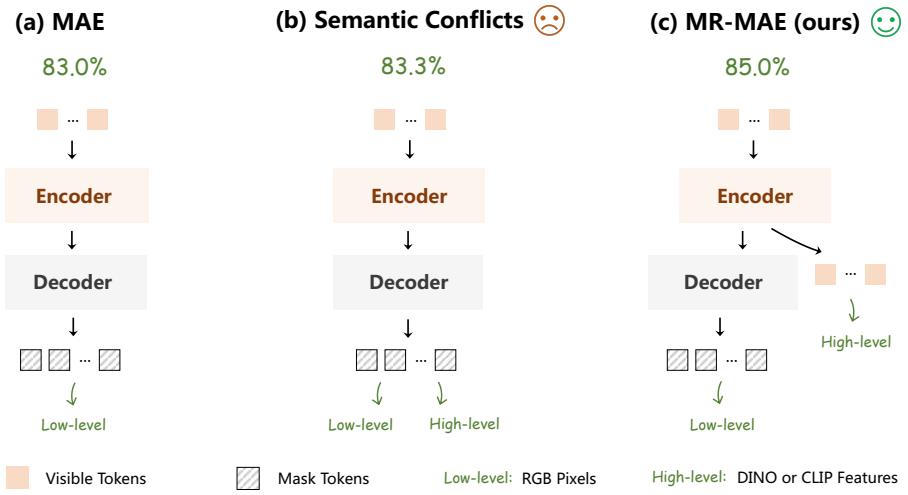

Figure 1: **Pre-training with MR-MAE.** (a) The original MAE only reconstructs low-level RGB pixels for masked tokens. (b) Applying both low-level and high-level supervisions to the decoder outputs causes semantic conflicts. (c) Our MR-MAE applies low-level and high-level supervisions respectively to different image tokens and network layers. The top-1 accuracy by fine-tuning on ImageNet-1K (Russakovsky et al., 2015) can be improved from 83.0% to 85.5%.

which are insufficient to guide the encoder and slows down the convergence speed of the pre-training stage.

To build more effective reconstruction targets, existing methods (Wei et al., 2022a; Baevski et al., 2022; Wei et al., 2022b; Peng et al., 2022; Hou et al., 2022) explore off-the-shelf pre-trained DINO (Caron et al., 2021), CLIP (Radford et al., 2021b), or online momentum features (He et al., 2020) as the high-level supervisions. However, considering the interference caused by simultaneous reconstruction of the two types of targets, previous methods simply replace the original RGB pixel targets by the high-level features and only use them to supervise the decoder's outputs.

Different from previous approaches (Wei et al., 2022a; Baevski et al., 2022; Wei et al., 2022b; Peng et al., 2022; Hou et al., 2022) that only apply high-level supervisions at the decoder, we aim to take advantages of both high-level semantics and low-level textures, and benefit the encoder's pre-training by learning with the two targets. To overcome the conflicts between two types of semantics, we introduce a new framework, named Mimic-before-Reconstruct Masked Autoencoders (MR-MAE). The original MAE randomly samples 25% visible tokens and processes them by the encoder. Then, the encoded tokens mixed with position embeddings are fed into the light-weight decoder for predicting pixel values of the 75% masked tokens. Our proposed MR-MAE augments the original MAE with a simple yet effective mimic loss (Hinton et al., 2015), which is applied to only the visible tokens directly after the encoder. The mimic loss minimizes the L2 distance between the MAE encoder's outputs and the high-level features generated from off-the-shelf pre-trained image-language (CLIP) (Radford et al., 2021b) or image-image (DINO) (Caron et al., 2021). Unlike the insufficiently supervised encoder in MAE, such mimic loss can provide effective and direct guidance on the encoder. As our mimic loss and the reconstruction loss are applied for different groups of tokens (25% visible vs 75% masked) and different network layers (encoder vs decoder's outputs), our MR-MAE well solves the supervision conflicts between the low-level and high-level learning targets. The learned representations even surpass the teacher networks (CLIP and DINO), demonstrating the benefit to jointly learn low-level and high-level targets.

Compared with the original MAE base model (He et al., 2022b) (83.8%) that aims to reconstruct low-level RGB pixels, our MR-MAE base model with a CLIP teacher not only enhances the ImageNet-1K fine-tuning accuracy to 85.0% (+1.2%), but also shortens the pre-training epochs from 1600 to 200. By further appending masked convolution stages introduced by MCMAE (Gao et al., 2022) and other tricks, our improved variant, MR-MCMAE base and huge-392 models, can attain 85.6% and 88.5% by merely pre-training for 200 epochs. This fully demonstrates the scaling ability of our approach. Notably, MR-MAE base and MR-MCMAE base surpass the ImageNet-1K (Rus-

sakovsky et al., 2015) fine-tuning accuracy of CLIP (84.2%) by +0.8% and +1.4%, respectively. This indicates that our MR-MAE learns even better representations than the teacher network, while the performance of traditional knowledge distillation is upper-bounded by the teacher.

2 Related Work

2.1 Contrastive Learning

Contrastive learning (Chen et al., 2020; Wu et al., 2018; Caron et al., 2021; Radford et al., 2021b) has achieved great successes on learning effective visual representations by extracting invariances from augmented views of a signal source. DINO (Caron et al., 2021) and CLIP (Radford et al., 2021b) are two canonical approaches among contrastive learning paradigms. DINO (Caron et al., 2021) observed strong objectness emerges from ViT pre-trained by image-image contrastive learning. On the other hand, CLIP (Radford et al., 2021b) demonstrated amazing zero-shot ability through image-text pair contrastive learning. Although DINO and CLIP exhibits strong objectness cues and open-world recognition ability, the fine-tuning performance on downstream tasks are inferior to representations learned through MAE (He et al., 2022b) manner. Our MR-MAE borrows the high-level semantics extracted from off-the-shelf DINO or CLIP to supervise the features of visible tokens in MAE. Thanks to the guidance of teacher networks, MR-MAE can significantly improve the representations of MAE and shorten the training epochs.

2.2 Masked Image Modeling

Pre-training on large-scale unsupervised corpus with Masked Language Modeling (MLM) (Devlin et al., 2018) have shown superior performance on natural language understanding and generation. Motivated by MLM, BEiT (Bao et al., 2021) explored Masked Image Modeling (MIM) on vision transformers by reconstructing the vision dictionary extracted with DALL-E Ramesh et al. (2021; 2022). MAE (He et al., 2022a) further proposed an asymmetric encoder and decoder for scaling up MIM to huge models. Besides, it demonstrated a simple pixel reconstruction loss can learn good visual representations. Due to the simplicity and computational efficiency, MAE is raising to a popular generative pre-training paradigm. As MAE reconstructs low-level signals with an isotropic vision transformer architecture, researchers improve MAE by exploring high-level signals and hierarchical architectures. MaskFeat (Wei et al., 2022a), data2vec (Baevski et al., 2022), MVP (Wei et al., 2022b) and MILAN (Hou et al., 2022) revealed various high-level signals, such as pre-trained DINO (Caron et al., 2021), HOG features (Dalal & Triggs, 2005), momentum features (He et al., 2020) and multi-modality features (Radford et al., 2021b), which are more effective than reconstructing low-level signals. Different from those approaches that explore high-level features as new reconstruction targets of masked regions, MR-MAE utilizes high-level features for regularizing the representations of visible tokens produced by MAE encoder. Thus, our MR-MAE can take advantages of both low-level and high-level information. FD (Wei et al., 2022c) proposed to improve pre-trained contrastive representations through feature distillation. Compared with FD to feed all tokens into the encoder, the encoder of MR-MAE only processes partially visible tokens (e.g., 25%) which leads to a significantly decrease of GPU memory. DMAE (Bai et al., 2022) proposed to jointly optimize the reconstruction loss and align the features with pre-trained MAE teacher. As the MAE teacher is still pre-trained by reconstructing low-level signals, the representations of DAME still lack high-level semantics. Different from DMAE, MR-MAE guides the feature distillation with contrastively pre-trained features which are complementary with low-level signals. MCMAE (Gao et al., 2022), UM-MAE (Li et al., 2022b), MixMIM (Liu et al., 2022) and GreenMIM (Huang et al., 2022a) explore efficient and effective MIM frameworks with hierarchical vision transformers (Liu et al., 2021; Gao et al., 2021; Li et al., 2022a; Xiao et al., 2021). Our improved variant, MR-MCMAE, leverages the masked convolution stages in MCMAE to hierarchically encode visual representations.

3 Method

3.1 Revisiting MAE

Masked Autoencoders (MAE) (He et al., 2022b) employ an asymmetric encoder-decoder design for computationally efficient masked image modeling. Given an input image, MAE first divides it into

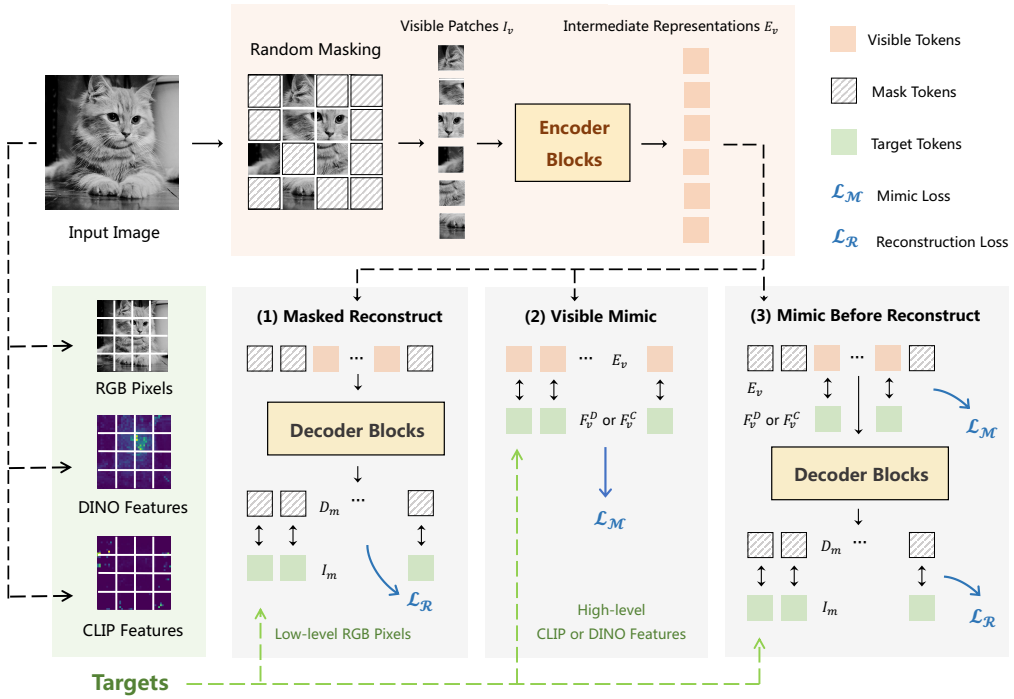

Figure 2: **Architecture of MR-MAE.** During MAE pre-training, we set both high-level and low-level learning targets respectively for different image tokens and network layers: mimic loss for 25% visible tokens of the encoder, and reconstruction loss for 75% masked tokens of the decoder.

patches of size $p \times p$, and randomly masks 75% of them. We denote the masked and visible patches respectively as $I_m \in \mathbb{R}^{l_m \times p^2}$ and $I_v \in \mathbb{R}^{l_v \times p^2}$, where l_m and l_v denote the numbers of masked and visible tokens. Then, the 25% visible patches are tokenized and fed into a transformer encoder to produce the C-dimensional intermediate representation $E_v \in \mathbb{R}^{l_v \times C}$. As shown in Figure 2 (1), MAE employs a light-weight transformer decoder to predict $D_m \in \mathbb{R}^{l_m \times p^2}$ to reconstruct RGB values of the masked tokens. An L2 reconstruction loss $\mathcal{L}_\mathcal{R}$ between D_m and I_m is used:

$$\mathcal{L}_\mathcal{R} = \frac{1}{l_m} \| D_m - I_m \|_2^2. \tag{1}$$

Despite its promising transfer capacity, MAE requires costly 1600 epochs to be fully pre-trained, which is partially due to the missing guidance to the intermediate representations of the encoder. Furthermore, by visualizing the attention map of [CLS] tokens in MAE's encoder as shown in Figure 5, we observe that MAE focuses more on some detailed texture patterns than the centric objects, since merely low-level RGB values I_m serve as the reconstruction targets. Therefore, we argue that the low-level supervision at the decoder's outputs not only slows down the pre-training convergence of MAE, but also limits its representations to capture high-level semantics.

3.2 Mimic before Reconstruct

To address the above issues, we propose to **M**imic before **R**econstruct for Masked Autoencoders, termed as MR-MAE, which is a simple and effective strategy to enhance MAE (He et al., 2022b) by regularizing the intermediate representations with pre-trained off-the-shelf feature encoders. The overall pipeline of MR-MAE is illustrated in Figure 2. Following MAE, MR-MAE also inputs the visible 25% tokens into the transformer encoder to encode the intermediate representation E_v. Different from only supervising the low-level reconstruction after the decoder, we propose to guide the intermediate E_v from the encoder with high-level features produced by DINO or CLIP, which contain rich high-level semantics, as shown in Figure 2 (2). We first extract the DINO or CLIP features by feeding the input image into their transformer-based visual encoders, denoted as

$F_v^D, F_v^C \in \mathbb{R}^{l_v \times C}$. By appending a feature mimic head on top of the encoder, we transform the visible representations E_v via a linear projection layer to mimic F_v^D or F_v^C. The L2 mimic loss of MR-MAE is defined as:

$$\mathcal{L}_{\mathcal{M}} = \frac{1}{l_v} \| \mathrm{L}(E_v) - F_v \|_2^2, \tag{2}$$

where F_v denotes either DINO's F_v^D or CLIP's F_v^C while L denotes mimic head.

To incorporate both low-level and high-level information, we also apply a light-weight decoder in MR-MAE to reconstruct the 75% masked RGB pixels, as shown in Figure 2 (3). We adopt the L2 reconstruction loss $\mathcal{L}_{\mathcal{R}}$ in Eq. 1 between D_m and I_m. As the feature mimic loss $\mathcal{L}_{\mathcal{M}}$ for visible tokens and the reconstruction loss $\mathcal{L}_{\mathcal{R}}$ for masked tokens aim at encoding different aspects of the input image, i.e., high-level semantics and low-level textures, they can complement each other to learn more discriminative representations. In addition, MR-MAE avoids the conflict of learning between low-level and high-level targets by applying supervisions upon different groups of tokens (25% visible vs 75% masked) and different network layers (encoder vs decoder's outputs). With the newly introduced high-level feature mimic loss, our proposed MR-MAE significantly improves the downstream performance of MAE and shortens its pre-training epochs.

3.3 Bag-of-tricks for MR-MAE

To further unleash the learning potential, we borrow some tricks from previous approaches and integrate them into MR-MAE to enhance our learned representations. We denote the improved variant as **MR-MCMAE**.

Focused Mimicking. MAE adopts a random masking strategy for visible token selection, which is a natural choice for low-level signal reconstruction without additional guidance. As the [CLS] token in off-the-shelf pre-trained models can clearly delineate regions of importance (Caron et al., 2021) via its attention map, we select the most salient tokens in teacher network's attention maps for visible feature mimicking. In this way, MR-MAE can better capture informative high-level semantics encoded in the teacher network, rather than the non-salient low-level ones. Similar strategies were previously discussed in MST (Li et al., 2021b), ADIOS (Shi et al., 2022b), AttnMASK (Kakogeorgiou et al., 2022), and MILAN (Hou et al., 2022).

Multi-layer Fusion. The original MAE only feeds the output tokens from the encoder's last layer into the decoder for masked pixel reconstruction. As different layers of the encoder might depict different abstraction levels of an image, we fuse the visible tokens from multiple intermediate layers of the encoder by element-wisely addition, and then utilize the fused ones for high-level feature mimicking and low-level pixel reconstruction. By this, the supervision from feature mimicking can be directly applied to multiple layers of the encoder, leading to the improved visual representations. Similar results have been demonstrated in BERT (Shi et al., 2022a), contrastive learning (Wang et al., 2022), and hierarchical MIM (Gao et al., 2022).

Masked Convolution Stages. Exploring multi-scale visual information has achieved great successes on computer vision tasks as objects exist in various scales. Following MCMAE (Gao et al., 2022), we append extra masked convolution stages before the transformer blocks (Gao et al., 2022; Xiao et al., 2021; Gao et al., 2021; Guo et al., 2022) to efficiently capture high-resolution details, and apply multi-scale block-wise masking (Gao et al., 2022) to prevent information leakage for pixel reconstruction. Such multi-scale encoding can learn hierarchical representations and achieve significant improvements on downstream tasks.

4 Experiments

For image classification, we pre-train our final model, MR-MCMAE (i.e., MR-MAE enhanced by the MCMAE architectures and other bug-of-tricks), on ImageNet-1K (Russakovsky et al., 2015), and compare with state-of-the-art Masked Image Modeling (MIM) methods by fine-tuning for top-1 accuracy. To further evaluate MR-MCMAE on high-resolution images, we fine-tune our pre-trained model on COCO (Lin et al., 2014) with Mask-RCNN (He et al., 2017) framework, and report AP^{box}

Table 1: Image classification by fine-tuning on ImageNet-1K (Russakovsky et al., 2015). 'Ratio' denotes the visible ratio of image tokens fed into the encoder. 'P-Epochs' and 'FT' denote pre-training epochs and the top-1 accuracy by fine-tuning.

Methods	Backbone	Params. (M)	Supervision	Ratio	P-Epochs	FT (%)
BEiT (Bao et al., 2021)	ViT-B	88	DALLE	100%	300	83.0
MAE (He et al., 2022b)	ViT-B	88	RGB	25%	1600	83.6
CAE (Chen et al., 2022)	ViT-B	88	RGB	25%	800	83.6
MaskFeat (Wei et al., 2022a)	ViT-B	88	HOG	100%	300	83.6
SimMIM (Xie et al., 2022)	Swin-B	88	RGB	100%	800	84.0
DMAE (Bai et al., 2022)	ViT-B	88	MAE	25%	100	84.0
data2vec (Baevski et al., 2022)	ViT-B	88	Momentum	100%	800	84.2
MVP (Wei et al., 2022b)	ViT-B	88	CLIP	100%	300	84.4
MCMAE (Gao et al., 2022)	CViT-B	88	RGB	25%	1600	85.0
MixMIM (Liu et al., 2022)	MixMIM-B	88	RGB	100%	600	85.1
CMAE (Huang et al., 2022b)	CViT-B	88	RGB	25%	1600	85.3
MILAN (Hou et al., 2022)	ViT-B	88	CLIP	25%	400	85.4
BEiT V2 (Peng et al., 2022)	ViT-B	88	CLIP	100%	1600	85.5
MR-MCMAE	CViT-B	88	CLIP	25%	400	85.8

and AP^{mask} results. Then, we conduct extensive ablation studies over each component of MR-MCMAE to validate their effectiveness.

4.1 IMAGENET-1K PRE-TRAINING AND FINE-TUNING

Experiment Setups. We adopt the fullly-fledged MR-MCMAE base model as default for comparison. We follow the protocol of pre-training and fine-tuning on ImageNet-1K as previous approaches. Specifically, MR-MCMAE base is pre-trained for 400 epochs with batch size 1,024 and weight decay 0.05. We adopt the AdamW (Loshchilov & Hutter, 2018) optimizer and the cosine learning rate scheduler with an maximum learning rate 1.5×10^{-4} and 80-epoch warming up. We utilize the mask ratio 25% and 8 decoder blocks following the practices in MAE (He et al., 2022b). The pre-training of MR-MCMAE jointly optimizes the reconstruction loss and mimic loss, whose weights are 0.5 and 0.5. By default, we choose ViT-B/16 pre-trained by CLIP (Radford et al., 2021a) as the high-level teacher. After the self-supervised pre-training, we transfer the pre-trained encoder as an initialization for fine-tuning on ImageNet-1K and report the top-1 accuracy on the validation set. The fine-tuning takes 100 epochs with 5-epoch warming up. We adopt the same batch size, optimizer, and weight decay as pre-training. The initial learning rate, layer-wise learning rate decay, and drop path rate are set to be 3×10^{-4}, 0.6 and 0.2, respectively.

Results on ImageNet-1K Finetuning. We compare our MR-MCMAE base model with previous state-of-the-art approaches of the similar model size on Table 1. BeiT (Bao et al., 2021), MAE (He et al., 2022b), CAE (Chen et al., 2022) have validated Masked Image Modeling (MIM) paradigm to be effective approaches for pre-training vision transformers. Due to their reconstruction of low-level pixels and the adoption of isotropic architectures, our MR-MCMAE can surpass the performance of those approaches by large margins (85.8% vs 83.0/83.6/83.6/84.0%). SimMIM (Xie et al., 2022), MCMAE (Gao et al., 2022) and MixMIM (Liu et al., 2022) introduce multi-scale features into MIM, resulting in improved fine-tuning accuracy compared with the isotropic architectures. As previous multi-scale approaches still reconstruct low-level signals, our MR-MCMAE can surpass their fine-tuning accuracy (85.8% vs 84.0/85.0/85.1%) with fewer pre-training epochs (400 vs 800/1600/600).

Another line of researches focuses on directly replacing the reconstruction of low-level signals with high-level semantic targets, MaskFeat (Wei et al., 2022a), data2vec (Baevski et al., 2022), MVP (Wei et al., 2022b) and MILAN (Hou et al., 2022) demonstrate promising results by integrating DINO (Zhang et al., 2022), momentum features (He et al., 2020) and CLIP (Radford et al., 2021a). MILAN (Hou et al., 2022) proposes a novel promoting decoder and semantic-aware masking to enhance the feature learning by reconstructing high-level features. BeiT V2 (Peng et al., 2022) replaces the original DALL-E tokenizers with high-level semantic tokenizers learned by self-encoding of CLIP features. Compared with advanced approaches for reconstructing high-level signals, such

Table 2: Object Detection by fine-tuning on COCO (Lin et al., 2014) based on the Mask-RCNN (He et al., 2017) framework. 'F-epochs' denotes the epochs for fine-tuning.

Methods	P-Epochs	F-Epochs	AP^{box}	AP^{mask}	Params. (M)	FLOPs (T)
ViTDet (Li et al., 2022c)	1600	100	51.2	45.5	111	0.8
CMAE (Huang et al., 2022b)	1600	25	52.9	47.0	104	0.9
MCMAE (Gao et al., 2022)	1600	25	53.2	47.1	104	0.9
MR-MCMAE	400	25	53.4	46.9	104	0.9

Table 3: Ablation study for 'mimic before reconstruct' and the bag-of-tricks for MR-MAE.

P-Epochs	Low Level	High Level	Focused Mimic	Multi-layer Fusion	Masked Conv.	ImageNet-1K FT	COCO AP^{box}	COCO AP^{mask}
200	$\mathcal{L}_{\mathcal{R}}$					83.0	N/A	N/A
	$\mathcal{L}_{\mathcal{R}}$	$\mathcal{L}_{\mathcal{M}}$				84.7	N/A	N/A
	$\mathcal{L}_{\mathcal{R}}$	$\mathcal{L}_{\mathcal{M}}$	✓			84.9	N/A	N/A
	$\mathcal{L}_{\mathcal{R}}$	$\mathcal{L}_{\mathcal{M}}$	✓	✓		85.0	51.6	45.5
	$\mathcal{L}_{\mathcal{R}}$	$\mathcal{L}_{\mathcal{R}}$	✓	✓		83.3	50.3	44.9
		$\mathcal{L}_{\mathcal{M}}$	✓	✓		84.9	50.9	44.8
	$\mathcal{L}_{\mathcal{R}}$	$\mathcal{L}_{\mathcal{M}}$	✓	✓	✓	85.5	53.0	46.5

as MILAN and BeiT-V2, MR-MCMAE still achieves better performance (85.8% vs 85.4/85.5%), since we jointly learn low-level and high-level targets with multi-scale architectures. CMAE (Huang et al., 2022b) learns representations through joint optimization of contrastive loss and reconstruction loss. Different from CMAE, MR-MCMAE utilizes a teacher model pre-trained from large-scale image-text contrastive learning, which contains more abundant semantic knowledge. MR-MCMAE improves the top-1 accuracy of CMAE from 85.3% to 85.8% and shortens the pre-training epochs from 1600 to 400. DMAE (Bai et al., 2022) adopts a similar approach as MR-MCMAE, which mimics features generated from the pre-trained teacher and reconstructs the low-level pixels. However, since the teacher of DMAE is still pre-trained with low-level pixel targets, the fine-tuning accuracy of DMAE is inferior to MR-MCMAE (84.0% vs 85.8%).

4.2 Object Detection

Experiment Setups. We evaluate the downstream transfer capacity of MR-MCMAE on the widely adopted COCO dataset (Lin et al., 2014). We apply the pre-trained encoder of MR-MCMAE as initialization of backbone for Mask-RCNN. Following ViTDet (Li et al., 2021a; 2022c), we simply expand the features for multiple scales as an alternative of feature pyramid network (FPN) (Lin et al., 2017). The resolution of the input image, learning rate, and layer decay are set as $1,024 \times 1,024$, 2×10^{-4} and 0.8, respectively. The model is fine-tuned for 25 epochs with batch size 16.

Results on COCO Fine-tuning. In Table 2, we use our proposed MR-MCMAE as the pre-trained backbone for Mask-RCNN (He et al., 2017). MR-MCMAE attains 53.4% AP^{box} and 46.9% AP^{mask} by fine-tuning 25 epochs on the COCO train2017 split. Compared with the baseline ViTDet (Li et al., 2022c), which adopts the encoder of MAE pre-trained for 1600 epochs, MR-MCMAE can improve AP^{box} and AP^{mask} by +2.2% and +1.4%. Besides, we shorten the pre-training epochs from 1600 to 400 and the fine-tuning epochs from 100 to 25. Compared with multi-scale backbones, such as CMAE (Huang et al., 2022b) and MCMAE (Gao et al., 2022), MR-MCMAE achieves comparable AP^{box} and AP^{mask} with a much shorter pre-training epochs (1600 vs 400 epochs).

4.3 Ablation Studies

To validate each component of MR-MCMAE, we conduct the following ablation studies.

Figure 3: Ablation study for the influence of pre-training epochs on ImageNet-1K and COCO object detection.

P-Epochs	ImageNet-1K FT	COCO AP^{box}	COCO AP^{mask}
200	85.5	52.7	44.8
400	85.8	53.4	46.9
800	85.8	53.5	47.0

Figure 4: Ablation study for the influence of high-level pre-training targets on the ImageNet-1K fine-tuning accuracy.

High-level Target	Params (M)	FT
DINO	88	84.0
CLIP	88	85.0
CLIP/DINO (Joint)	88	83.8
CLIP/DINO (Sep.)	176	85.5

Mimic Before Reconstruct. As shown in the first row of Table 3, the baseline MAE model with the low-level reconstruction loss achieves 83.0% fine-tuning accuracy on ImageNet-1K with 200-epoch pre-training. By jointly learning with the mimic loss, the classification accuracy is boosted by +1.7%. The comparison between the forth and sixth rows of Table 3 indicates that the joint optimization of both low-level and high-level targets can achieve better performance than only mimicking high-level semantics, especially for AP^{box} of object detection (+0.7%).

Bag-of-tricks. In Table 3, we also ablate each trick mentioned in Section 3.3. Based on the 84.7% fine-tuning accuracy with both $\mathcal{L}_{\mathcal{R}}$ and $\mathcal{L}_{\mathcal{M}}$, Focused mimicking leads to +0.2% improvement due to the focus of salient tokens guided by attention maps of the teacher network. Multi-layer Fusion further improves the accuracy by +0.1%. The introduction of Masked Convolution Stages increases the ImageNet-1K fine-tuning accuracy by +0.5%. More importantly, it improves AP^{box} and AP^{mask} by +1.4% and +1.0%, respectively, demonstrating the significance of multi-scale architectures.

Conflicts between Low-level and High-level Targets. As low-level and high-level targets contain different visual semantics, their joint supervisions might conflict with each other. As shown in the forth and fifth rows of Table 3, joint reconstruction of low-level and high-level targets deteriorates ImageNet-1K fine-tuning accuracy by -1.7%, AP^{box} by -1.3% and AP^{mask} by -0.6%. The results indicate our Mimic-before-Reconstruct framework is able to solve the conflicts between low-level and high-level targets by applying mimic and reconstruction losses upon different groups of tokens (visible vs masked) and different network layers (encoder vs decoder's outputs).

Different High-level Targets. As image-image contrastive learning (DINO) and image-language contrastive learning (CLIP) encode different high-level semantics. We ablate the performance of MR-MAE base with different high-level semantics. As shown in Table 4, features generated by CLIP can surpass DINO by +1%. This implies image-language contrastive learning provides stronger high-level semantics than image-image contrasive learning. The joint mimicking of multiple high-level signals is worse than independent mimicking. We hypothesize that the performance degradation is due to the gradient conflicts of predicting different high-level targets. To avoid the degradation introduced by the conflicts of reconstructing different high-level targets, we separately pre-train and fine-tune MR-MAE with different high-level targets then ensemble the two models. As shown in Table 4, CLIP/DINO (Sep.) can surpass the CLIP/DINO (Joint) by +1.7%, which validates the complementary representation learned with different targets. In the future, we will explore more efficient approaches to better incorporate multiple pre-trained high-level signals into a single student network.

Longer Pre-training Epochs. We ablate the influence of pre-training epochs on MR-MCMAE in Table 3. MR-MCMAE pre-trained for 200 epochs can achieve 85.5% ImageNet-1K fine-tuning accuracy and 52.7% AP^{box} for COCO. MR-MCMAE pre-trained for 400 epochs can improve ImageNet-1K fine-tuning accuracy by +0.3% and AP^{box} by +0.7%. Given longer pre-training epochs, such as 800 epochs, the performance saturates as shown in Table 3. This implies the introduction of high-level targets can make MIM approach converge much faster. The previous prolonged 1600 pre-training schedule can be shorted to 400 epochs under our Mimic-before-reconstruct framework.

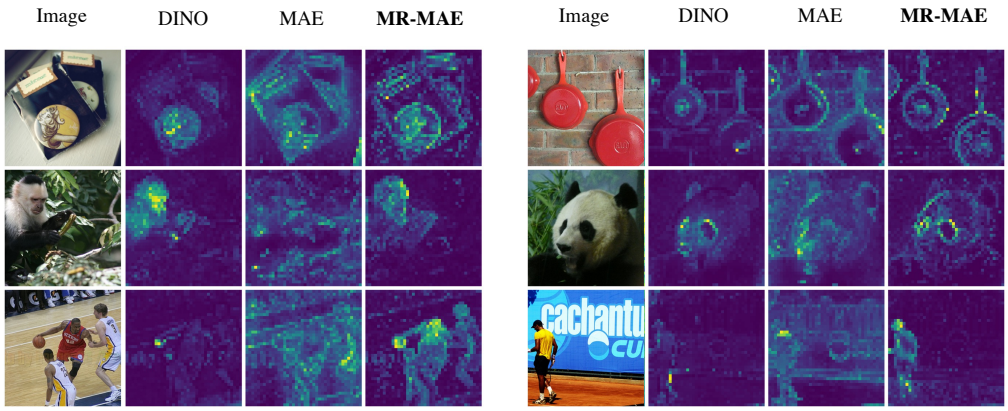

Figure 5: Visualization of attention weights at the last self-attention layer in DINO (Zhang et al., 2022), MAE (He et al., 2022a), and **MR-MAE (ours)**. MR-MAE can better capture salient feature representation compared to previous methods.

Table 4: ImageNet-1K finetuning accuracy of different model scales.

Method	P-Epochs	Small	Base	Large	Huge	Huge-393	Huge-448
MAE (He et al., 2022a)	1600	79.5	83.6	85.9	86.9	-	87.8
MCMAE (Gao et al., 2022)	800	82.6	84.6	84.9	86.2	-	-
MR-MCMAE	200	83.6	85.5	86.8	88.0	88.5	-

Scaling-up the Model. To test the scalability of our framework, we experiment with different models size of MR-MCMAE and reported the ImageNet-1K fine-tuning accuracy on Table 4. Compared with the single-scale baseline MAE and the stronger multi-scale baseline MCMAE, our MR-MCMAE demonstrates significantly improved performance over all model sizes with much shortened pre-training epochs.

Feature Visualization. To provide intuitions on why high-level targets improve the representation, we visualize the attention map of [CLS] token of the last self-attention layer of different models. As shown in Figure 5, the attention of MAE is biased towards texture patterns due to its aim of low-level pixel reconstruction, implying that MAE waste its capacity on low-level textures irrelevant for semantic understanding. On the other side, the attention of DINO's [CLS] token overemphasises on partial information of salient object. The attention of our MR-MCMAE can capture complete object information compared with DINO and MAE.

5 Conclusion

In this paper, we propose MR-MAE, a simple and effective framework for masked image modeling, which conducts feature mimicking before pixel reconstruction to incorporate high-level semantics into MAE. Specifically, for the 25% visible tokens from the encoder, we apply a mimic loss upon them to learn the semantic information encoded by off-the-shelf pre-trained models. For the 75% masked tokens after the decoder, we preserve the original reconstruction loss to model low-level texture patterns. By this, our MR-MAE does not only model both high-level and low-level information, but also well solves the semantic conflicts between the two types of targets. Furthermore, our variant MR-MCMAE built with bag-of-tricks can achieve superior performance for image classification and downstream detection.

Limitation: Although MR-MAE effectively learns the high-level knowledge from CLIP or DINO, naive joint supervision of CLIP and DINO cannot achieve higher results (separate supervision first and model ensemble later can improve). Our future direction will focus on how to better guide MAE by high-level semantics from multiple teacher networks.

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. 2, 3, 6

Yutong Bai, Zeyu Wang, Junfei Xiao, Chen Wei, Huiyu Wang, Alan Yuille, Yuyin Zhou, and Cihang Xie. Masked autoencoders enable efficient knowledge distillers. *arXiv preprint arXiv:2208.12256*, 2022. 3, 6, 7

Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021. 3, 6

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 1

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9650–9660, 2021. 1, 2, 3, 5

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020. 3

Xiaokang Chen, Mingyu Ding, Xiaodi Wang, Ying Xin, Shentong Mo, Yunhao Wang, Shumin Han, Ping Luo, Gang Zeng, and Jingdong Wang. Context autoencoder for self-supervised representation learning. *arXiv preprint arXiv:2202.03026*, 2022. 6

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. 3

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. 1, 3

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. 1

Peng Gao, Jiasen Lu, Hongsheng Li, Roozbeh Mottaghi, and Aniruddha Kembhavi. Container: Context aggregation network. *arXiv preprint arXiv:2106.01401*, 2021. 3, 5

Peng Gao, Teli Ma, Hongsheng Li, Jifeng Dai, and Yu Qiao. Convmae: Masked convolution meets masked autoencoders. *arXiv preprint arXiv:2205.03892*, 2022. 2, 3, 5, 6, 7, 9

Jianyuan Guo, Kai Han, Han Wu, Yehui Tang, Xinghao Chen, Yunhe Wang, and Chang Xu. Cmt: Convolutional neural networks meet vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12175–12185, 2022. 5

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969, 2017. 5, 7

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020. 1, 2, 3, 6

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. 2022a. 3, 9

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16000–16009, 2022b. 1, 2, 3, 4, 6

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

Zejiang Hou, Fei Sun, Yen-Kuang Chen, Yuan Xie, and Sun-Yuan Kung. Milan: Masked image pretraining on language assisted representation. *arXiv preprint arXiv:2208.06049*, 2022. 2, 3, 5, 6

Lang Huang, Shan You, Mingkai Zheng, Fei Wang, Chen Qian, and Toshihiko Yamasaki. Green hierarchical vision transformer for masked image modeling. *arXiv preprint arXiv:2205.13515*, 2022a. 3

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*, 2022b. 6, 7

Ioannis Kakogeorgiou, Spyros Gidaris, Bill Psomas, Yannis Avrithis, Andrei Bursuc, Konstantinos Karantzalos, and Nikos Komodakis. What to hide from your students: Attention-guided masked image modeling. *arXiv preprint arXiv:2203.12719*, 2022. 5

Kunchang Li, Yali Wang, Junhao Zhang, Peng Gao, Guanglu Song, Yu Liu, Hongsheng Li, and Yu Qiao. Uniformer: Unifying convolution and self-attention for visual recognition. *arXiv preprint arXiv:2201.09450*, 2022a. 3

Xiang Li, Wenhai Wang, Lingfeng Yang, and Jian Yang. Uniform masking: Enabling mae pre-training for pyramid-based vision transformers with locality. *arXiv preprint arXiv:2205.10063*, 2022b. 3

Yanghao Li, Saining Xie, Xinlei Chen, Piotr Dollar, Kaiming He, and Ross Girshick. Benchmarking detection transfer learning with vision transformers. *arXiv preprint arXiv:2111.11429*, 2021a. 7

Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. *arXiv preprint arXiv:2203.16527*, 2022c. 7

Zhaowen Li, Zhiyang Chen, Fan Yang, Wei Li, Yousong Zhu, Chaoyang Zhao, Rui Deng, Liwei Wu, Rui Zhao, Ming Tang, et al. Mst: Masked self-supervised transformer for visual representation. *Advances in Neural Information Processing Systems*, 34:13165–13176, 2021b. 5

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*, pp. 740–755. Springer, 2014. 5, 7

Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2117–2125, 2017. 7

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*, 2022. 3, 6

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10012–10022, 2021. 3

Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam. 2018. 6

Zhiliang Peng, Li Dong, Hangbo Bao, Qixiang Ye, and Furu Wei. Beit v2: Masked image modeling with vector-quantized visual tokenizers. *arXiv preprint arXiv:2208.06366*, 2022. 2, 6

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 1

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021a. 6

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021b. 1, 2, 3

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*, pp. 8821–8831. PMLR, 2021. 3

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 3

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015. 2, 5, 6

Han Shi, Jiahui Gao, Hang Xu, Xiaodan Liang, Zhenguo Li, Lingpeng Kong, Stephen Lee, and James T Kwok. Revisiting over-smoothing in bert from the perspective of graph. *arXiv preprint arXiv:2202.08625*, 2022a. 5

Yuge Shi, N Siddharth, Philip Torr, and Adam R Kosiorek. Adversarial masking for self-supervised learning. In *International Conference on Machine Learning*, pp. 20026–20040. PMLR, 2022b. 5

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*, 2022. 5

Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14668–14678, 2022a. 2, 3, 6

Longhui Wei, Lingxi Xie, Wengang Zhou, Houqiang Li, and Qi Tian. Mvp: Multimodality-guided visual pre-training. *arXiv preprint arXiv:2203.05175*, 2022b. 2, 3, 6

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. 3

Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3733–3742, 2018. 3

Tete Xiao, Mannat Singh, Eric Mintun, Trevor Darrell, Piotr Dollár, and Ross Girshick. Early convolutions help transformers see better. *Advances in Neural Information Processing Systems*, 34:30392–30400, 2021. 3, 5

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9653–9663, 2022. 6

Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022. 6, 9