This paper shows that masked autoencoders (MAE) are scalable self-supervised learners for computer vision. Our MAE approach is simple: we mask random patches of the input image and reconstruct the missing pixels. It is based on two core designs. First, we develop an asymmetric encoder-decoder architecture, with an encoder that operates only on the visible subset of patches (without mask tokens), along with a lightweight decoder that reconstructs the original image from the latent representation and mask tokens. Second, we find that masking a high proportion of the input image, e.g., 75%, yields a nontrivial and meaningful self-supervisory task. Coupling these two designs enables us to train large models efficiently and effectively: we accelerate training (by $3 \times$ or more) and improve accuracy. Our scalable approach allows for learning high-capacity models that generalize well: e.g., a vanilla ViT-Huge model achieves the best accuracy (87.8%) among methods that use only ImageNet-1K data. Transfer performance in downstream tasks outperforms supervised pre-training and shows promising scaling behavior.

1. Introduction

Deep learning has witnessed an explosion of architectures of continuously growing capability and capacity [33, 25, 57]. Aided by the rapid gains in hardware, models today can easily overfit one million images [13] and begin to demand hundreds of millions of—often publicly inaccessible—labeled images [16].

This appetite for data has been successfully addressed in natural language processing (NLP) by self-supervised pre-training. The solutions, based on autoregressive language modeling in GPT [47, 48, 4] and masked autoencoding in BERT [14], are conceptually simple: they remove a portion of the data and learn to predict the removed content. These methods now enable training of generalizable NLP models containing over one hundred billion parameters [4].

The idea of masked autoencoders, a form of more general denoising autoencoders [58], is natural and applicable in computer vision as well. Indeed, closely related research in vision [59, 46] preceded BERT. However, despite significant interest in this idea following the success of BERT, progress of autoencoding methods in vision lags behind NLP. We ask: what makes masked autoencoding different between vision and language? We attempt to answer this question from the following perspectives:

(i) Until recently, architectures were different. In vision, convolutional networks [34] were dominant over the last decade [33]. Convolutions typically operate on regular grids and it is not straightforward to integrate ‘indicators’ such as mask tokens [14] or positional embeddings [57] into convolutional networks. This architectural gap, however, has been addressed with the introduction of Vision Transformers (ViT) [16] and should no longer present an obstacle.

(ii) Information density is different between language and vision. Languages are human-generated signals that are highly semantic and information-dense. When training a model to predict only a few missing words per sentence, this task appears to induce sophisticated language understanding. Images, on the contrary, are natural signals with heavy spatial redundancy—e.g., a missing patch can be recovered from neighboring patches with little high-level un-
understanding of parts, objects, and scenes. To overcome this
difference and encourage learning useful features, we show
that a simple strategy works well in computer vision: mask-
ing a very high portion of random patches. This strategy
largely reduces redundancy and creates a challenging self-
supervisory task that requires holistic understanding beyond
low-level image statistics. To get a qualitative sense of our
reconstruction task, see Figures 2 – 4.

(iii) The autoencoder’s decoder, which maps the latent
representation back to the input, plays a different role be-
tween reconstructing text and images. In vision, the decoder
reconstructs pixels, hence its output is of a lower semantic
level than common recognition tasks. This is in contrast
to language, where the decoder predicts missing words that
contain rich semantic information. While in BERT the de-
coder can be trivial (an MLP) [14], we found that for im-
ages, the decoder design plays a key role in determining the
semantic level of the learned latent representations.

Driven by this analysis, we present a simple, effective,
and scalable form of a masked autoencoder (MAE) for
visual representation learning. Our MAE masks random
patches from the input image and reconstructs the missing
patches in the pixel space. It has an asymmetric encoder-
decoder design. Our encoder operates only on the visible
subset of patches (without mask tokens), and our decoder is
lightweight and reconstructs the input from the latent rep-
resentation along with mask tokens (Figure 1). Shifting
the mask tokens to the small decoder in our asymmetric
encoder-decoder results in a large reduction in computation.
Under this design, a very high masking ratio (e.g., 75%) can
achieve a win-win scenario: it optimizes accuracy while al-
lowing the encoder to process only a small portion (e.g.,
25%) of patches. This can reduce overall pre-training time
by 3× or more and likewise reduce memory consumption,
enabling us to easily scale our MAE to large models.

Our MAE learns very high-capacity models that gen-
eralize well. With MAE pre-training, we can train data-
hungry models like ViT-Large/-Huge [16] on ImageNet-1K
with improved generalization performance. With a vanilla
ViT-Huge model, we achieve 87.8% accuracy when fine-
tuned on ImageNet-1K. This outperforms all previous re-
sults that use only ImageNet-1K data. We also evaluate
transfer learning on object detection, instance segmentation,
and semantic segmentation. In these tasks, our pre-training
achieves better results than its supervised pre-training coun-
terparts, and more importantly, we observe significant gains
by scaling up models. These observations are aligned
with those witnessed in self-supervised pre-training in NLP
[14, 47, 48, 4] and we hope that they will enable our field to
explore a similar trajectory.
2. Related Work

**Masked language modeling** and its autoregressive counterparts, *e.g.*, BERT [14] and GPT [47, 48, 4], are highly successful methods for pre-training in NLP. These methods hold out a portion of the input sequence and train models to predict the missing content. These methods have been shown to scale excellently [4] and a large abundance of evidence indicates that these pre-trained representations generalize well to various downstream tasks.

**Autoencoding** is a classical method for learning representations. It has an encoder that maps an input to a latent representation and a decoder that reconstructs the input. For example, PCA and k-means are autoencoders [29]. Denoising autoencoders (DAE) [58] are a class of autoencoders that corrupt an input signal and learn to reconstruct the original, uncorrupted signal. A series of methods can be thought of as a generalized DAE under different corruptions, *e.g.*, masking pixels [59, 46, 6] or removing color channels [70]. Our MAE is a form of denoising autoencoding, but different from the classical DAE in numerous ways.


**Self-supervised learning** approaches have seen significant interest in computer vision, often focusing on different pretext tasks for pre-training [15, 61, 42, 70, 45, 17]. Recently, contrastive learning [3, 22] has been popular, *e.g.*, [62, 43, 23, 7], which models image similarity and dissimilarity (or only similarity [21, 8]) between two or more views. Contrastive and related methods strongly depend on data augmentation [7, 21, 8]. Autoencoding pursues a conceptually different direction, and it exhibits different behaviors as we will present.

3. Approach

Our masked autoencoder (MAE) is a simple autoencoding approach that reconstructs the original signal given its partial observation. Like all autoencoders, our approach has an encoder that maps the observed signal to a latent representation, and a decoder that reconstructs the original signal from the latent representation. Unlike classical autoencoders, we adopt an asymmetric design that allows the encoder to operate only on the partial, observed signal (without mask tokens) and a lightweight decoder that reconstructs the full signal from the latent representation and mask tokens. Figure 1 illustrates the idea, introduced next.

**Masking.** Following ViT [16], we divide an image into regular non-overlapping patches. Then we sample a subset of patches and mask (*i.e.*, remove) the remaining ones. Our sampling strategy is straightforward: we sample random patches without replacement, following a uniform distribution. We simply refer to this as “random sampling”.

Random sampling with a high masking ratio (*i.e.*, the ratio of removed patches) largely eliminates redundancy, thus creating a task that cannot be easily solved by extrapolation from visible neighboring patches (see Figures 2–4). The uniform distribution prevents a potential center bias (*i.e.*, more masked patches near the image center). Finally, the highly sparse input creates an opportunity for designing an efficient encoder, introduced next.

**MAE encoder.** Our encoder is a ViT [16] but applied only on visible, unmasked patches. Just as in a standard ViT, our encoder embeds patches by a linear projection with added positional embeddings, and then processes the resulting set via a series of Transformer blocks. However, our encoder only operates on a small subset (*e.g.*, 25%) of the full set. Masked patches are removed; no mask tokens are used. This allows us to train very large encoders with only a fraction of compute and memory. The full set is handled by a lightweight decoder, described next.

**MAE decoder.** The input to the MAE decoder is the full set of tokens consisting of (i) encoded visible patches, and (ii) mask tokens. See Figure 1. Each mask token [14] is a shared, learned vector that indicates the presence of a miss-
ing patch to be predicted. We add positional embeddings to all tokens in this full set; without this, mask tokens would have no information about their location in the image. The decoder has another series of Transformer blocks.

The MAE decoder is only used during pre-training to perform the image reconstruction task (only the encoder is used to produce image representations for recognition). Therefore, the decoder architecture can be flexibly designed in a manner that is independent of the encoder design. We experiment with very small decoders, narrower and shallower than the encoder. For example, our default decoder has <10% computation per token vs. the encoder. With this asymmetrical design, the full set of tokens are only processed by the lightweight decoder, which significantly reduces pre-training time.

**Reconstruction target.** Our MAE reconstructs the input by predicting the pixel values for each masked patch. Each element in the decoder’s output is a vector of pixel values representing a patch. The last layer of the decoder is a linear projection whose number of output channels equals the number of pixel values in a patch. The decoder’s output is reshaped to form a reconstructed image. Our loss function computes the mean squared error (MSE) between the reconstructed and original images in the pixel space. We compute the loss only on masked patches, similar to BERT [14].

We also study a variant whose reconstruction target is the normalized pixel values of each masked patch. Specifically, we compute the mean and standard deviation of all pixels in a patch and use them to normalize this patch. Using normalized pixels as the reconstruction target improves representation quality in our experiments.

**Simple implementation.** Our MAE pre-training can be implemented efficiently, and importantly, does not require any specialized sparse operations. First we generate a token for every input patch (by linear projection with an added positional embedding). Next we randomly shuffle the list of tokens and remove the last portion of the list, based on the masking ratio. This process produces a small subset of tokens for the decoder and is equivalent to sampling patches without replacement. After encoding, we append a list of mask tokens to the list of encoded patches, and unshuffle this full list (inverting the random shuffle operation) to align all tokens with their targets. The decoder is applied to this full list (with positional embeddings added). As noted, no sparse operations are needed. This simple implementation introduces negligible overhead as the shuffling and unshuffling operations are fast.

---

1Computing the loss only on masked patches differs from traditional denoising autoencoders [58] that compute the loss on all pixels. This choice is purely result-driven: computing the loss on all pixels leads to a slight decrease in accuracy (e.g., ~0.5%).

---

Figure 5. Masking ratio. A high masking ratio (75%) works well for both fine-tuning (top) and linear probing (bottom). The y-axes are ImageNet-1K validation accuracy (%) in all plots in this paper.

### 4. ImageNet Experiments

We do self-supervised pre-training on the ImageNet-1K (IN1K) [13] training set. Then we do supervised training to evaluate the representations with (i) end-to-end fine-tuning or (ii) linear probing. We report top-1 validation accuracy of a single $224 \times 224$ crop. Details are in Appendix A.1.

**Baseline: ViT-Large.** We use ViT-Large (ViT-L/16) [16] as the backbone in our ablation study. ViT-L is very big (an order of magnitude bigger than ResNet-50 [25]) and tends to overfit. The following is a comparison between ViT-L trained from scratch vs. fine-tuned from our baseline MAE:

<table>
<thead>
<tr>
<th></th>
<th>scratch, original [16]</th>
<th>scratch, our impl.</th>
<th>baseline MAE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>76.5</td>
<td>82.5</td>
<td>84.9</td>
</tr>
</tbody>
</table>

We note that it is nontrivial to train supervised ViT-L from scratch and a good recipe with strong regularization is needed (82.5%, see Appendix A.2). Even so, our MAE pretraining contributes a big improvement. Here fine-tuning is only for 50 epochs (vs. 200 from scratch), implying that the fine-tuning accuracy heavily depends on pre-training.

### 4.1. Main Properties

We ablate our MAE using the default settings in Table 1 (see caption). Several intriguing properties are observed.

**Masking ratio.** Figure 5 shows the influence of the masking ratio. The optimal ratios are surprisingly high. The ratio of 75% is good for both linear probing and fine-tuning. This behavior is in contrast with BERT [14], whose typical masking ratio is 15%. Our masking ratios are also much higher than those in related works [6, 16, 2] in computer vision (20% to 50%).

The model infers missing patches to produce different, yet plausible, outputs (Figure 4). It makes sense of the gestalt of objects and scenes, which cannot be simply completed by extending lines or textures. We hypothesize that this reasoning-like behavior is linked to the learning of useful representations.

Figure 5 also shows that linear probing and fine-tuning results follow different trends. For linear probing, the ac-
blocks ft lin
1 84.8 65.5
2 84.9 70.0
4 84.9 71.9
8 84.9 73.5
12 84.4 73.3

Decoder depth. A deep decoder can improve linear probing accuracy.

Decoder design. Our MAE decoder can be flexibly designed, as studied in Table 1a and 1b.

Table 1a varies the decoder depth (number of Transformer blocks). A sufficiently deep decoder is important for linear probing. This can be explained by the gap between a pixel reconstruction task and a recognition task: the last several layers in an autoencoder are more specialized for reconstruction, but are less relevant for recognition. A reasonably deep decoder can account for the reconstruction specialization, leaving the latent representations at a more abstract level. This design can yield up to 8% improvement in linear probing (Table 1a, ‘lin’). However, if fine-tuning is used, the last layers of the encoder can be tuned to adapt to the recognition task. The decoder depth is less influential for improving fine-tuning (Table 1a, ‘ft’).

Interestingly, our MAE with a single-block decoder can perform strongly with fine-tuning (84.8%). Note that a single Transformer block is the minimal requirement to propagate information from visible tokens to mask tokens. Such a small decoder can further speed up training.

In Table 1b we study the decoder width (number of channels). We use 512-d by default, which performs well under fine-tuning and linear probing. A narrower decoder also works well with fine-tuning.

Overall, our default MAE decoder is lightweight. It has 8 blocks and a width of 512-d (gray in Table 1). It only has 9% FLOPs per token vs. ViT-L (24 blocks, 1024-d). As such, while the decoder processes all tokens, it is still a small fraction of the overall compute.

Table 1. MAE ablation experiments with ViT-L/16 on ImageNet-1K. We report fine-tuning (ft) and linear probing (lin) accuracy (%). If not specified, the default is: the decoder has depth 8 and width 512, the reconstruction target is unnormalized pixels, the data augmentation is random resized cropping, the masking ratio is 75%, and the pre-training length is 800 epochs. Default settings are marked in gray.

<table>
<thead>
<tr>
<th>case</th>
<th>ft</th>
<th>lin</th>
</tr>
</thead>
<tbody>
<tr>
<td>pixel (w/o norm)</td>
<td>84.9</td>
<td>73.5</td>
</tr>
<tr>
<td>pixel (w/ norm)</td>
<td>85.4</td>
<td>73.9</td>
</tr>
<tr>
<td>PCA</td>
<td>84.6</td>
<td>72.3</td>
</tr>
<tr>
<td>dVAE token</td>
<td>85.3</td>
<td>71.6</td>
</tr>
</tbody>
</table>

Reconstruction target. Pixels as reconstruction targets are effective.

<table>
<thead>
<tr>
<th>dim</th>
<th>ft</th>
<th>lin</th>
</tr>
</thead>
<tbody>
<tr>
<td>128</td>
<td>84.9</td>
<td>69.1</td>
</tr>
<tr>
<td>256</td>
<td>84.8</td>
<td>71.3</td>
</tr>
<tr>
<td>512</td>
<td>84.9</td>
<td>73.5</td>
</tr>
<tr>
<td>768</td>
<td>84.4</td>
<td>73.1</td>
</tr>
<tr>
<td>1024</td>
<td>84.3</td>
<td>73.1</td>
</tr>
</tbody>
</table>

Data augmentation. Our MAE works with minimal or no augmentation.

<table>
<thead>
<tr>
<th>case</th>
<th>ft</th>
<th>lin</th>
</tr>
</thead>
<tbody>
<tr>
<td>random</td>
<td>75</td>
<td>84.9</td>
</tr>
<tr>
<td>block</td>
<td>50</td>
<td>83.9</td>
</tr>
<tr>
<td>grid</td>
<td>75</td>
<td>84.0</td>
</tr>
</tbody>
</table>

Mask sampling. Random sampling works the best. See Figure 6 for visualizations.

Table 2. Wall-clock time of our MAE training (800 epochs), benchmarked in 128 TPU-v3 cores with TensorFlow. The speedup is relative to the entry whose encoder has mask tokens (gray). The decoder width is 512, and the mask ratio is 75%. †: This entry is estimated by training ten epochs.

<table>
<thead>
<tr>
<th>encoder</th>
<th>dec. depth</th>
<th>ft acc</th>
<th>hours</th>
<th>speedup</th>
</tr>
</thead>
<tbody>
<tr>
<td>ViT-L, w/ [M]</td>
<td>8</td>
<td>84.2</td>
<td>42.4</td>
<td>-</td>
</tr>
<tr>
<td>ViT-L</td>
<td>8</td>
<td>84.9</td>
<td>15.4</td>
<td>2.8×</td>
</tr>
<tr>
<td>ViT-L</td>
<td>1</td>
<td>84.8</td>
<td>11.6</td>
<td>3.7×</td>
</tr>
<tr>
<td>ViT-H, w/ [M]</td>
<td>8</td>
<td>-</td>
<td>119.6†</td>
<td>-</td>
</tr>
<tr>
<td>ViT-H</td>
<td>8</td>
<td>85.8</td>
<td>34.5</td>
<td>3.5×</td>
</tr>
<tr>
<td>ViT-H</td>
<td>1</td>
<td>85.9</td>
<td>29.3</td>
<td>4.1×</td>
</tr>
</tbody>
</table>

Mask token. An important design of our MAE is to skip the mask token [M] in the encoder and apply it later in the lightweight decoder. Table 1c studies this design.

If the encoder uses mask tokens, it performs worse: its accuracy drops by 14% in linear probing. In this case, there is a gap between pre-training and deploying: this encoder has a large portion of mask tokens in its input in pre-training, which does not exist in uncorrupted images. This gap may degrade accuracy in deployment. By removing the mask token from the encoder, we constrain the encoder to always see real patches and thus improve accuracy.

Moreover, by skipping the mask token in the encoder, we greatly reduce training computation. In Table 1c, we reduce the overall training FLOPs by 3.3×. This leads to a 2.8× wall-clock speedup in our implementation (see Table 2). The wall-clock speedup is even bigger (3.5–4.1×), for a smaller decoder (1-block), a larger encoder (ViT-H), or both. Note that the speedup can be >4× for a masking ratio of 75%, partially because the self-attention complexity is quadratic. In addition, memory is greatly reduced, which can enable training even larger models or speeding up more by large-batch training. The time and memory efficiency makes our MAE favorable for training very large models.
Reconstruction target. We compare different reconstruction targets in Table 1d. Our results thus far are based on pixels without (per-patch) normalization. Using pixels with normalization improves accuracy. This per-patch normalization enhances the contrast locally. In another variant, we perform PCA in the patch space and use the largest PCA coefficients (96 here) as the target. Doing so degrades accuracy. Both experiments suggest that the high-frequency components are useful in our method.

We also compare an MAE variant that predicts tokens, the target used in BEiT [2]. Specifically for this variant, we use the DALL-E pre-trained dVAE [50] as the tokenizer, following [2]. Here the MAE decoder predicts the token indices using cross-entropy loss. This tokenization improves fine-tuning accuracy by 0.4% vs. unnormalized pixels, but has no advantage vs. normalized pixels. It also reduces linear probing accuracy. In §5 we further show that tokenization is not necessary in transfer learning.

Our pixel-based MAE is much simpler than tokenization. The dVAE tokenizer requires one more pre-training stage, which may depend on extra data (250M images [50]). The dVAE encoder is a large convolutional network (40% FLOPs of ViT-L) and adds nontrivial overhead. Using pixels does not suffer from these problems.

Data augmentation. Table 1e studies the influence of data augmentation on our MAE pre-training.

Our MAE works well using cropping-only augmentation, either fixed-size or random-size (both having random horizontal flipping). Adding color jittering degrades the results and so we do not use it in other experiments.

Surprisingly, our MAE behaves decently even if using no data augmentation (only center-crop, no flipping). This property is dramatically different from contrastive learning and related methods [62, 23, 7, 21], which heavily rely on data augmentation. It was observed [21] that using cropping-only augmentation reduces the accuracy by 13% and 28% respectively for BYOL [21] and SimCLR [7]. In addition, there is no evidence that contrastive learning can work without augmentation: the two views of an image are the same and can easily satisfy a trivial solution.

In MAE, the role of data augmentation is mainly performed by random masking (ablated next). The masks are different for each iteration and so they generate new training samples regardless of data augmentation. The pretext task is made difficult by masking and requires less augmentation to regularize training.

Mask sampling strategy. In Table 1f we compare different mask sampling strategies, illustrated in Figure 6.

The block-wise masking strategy, proposed in [2], tends to remove large blocks (Figure 6 middle). Our MAE with block-wise masking works reasonably well at a ratio of 50%, but degrades at a ratio of 75%. This task is harder than that of random sampling, as a higher training loss is observed. The reconstruction is also blurrier.

We also study grid-wise sampling, which regularly keeps one of every four patches (Figure 6 right). This is an easier task and has lower training loss. The reconstruction is sharper. However, the representation quality is lower.

Simple random sampling works the best for our MAE. It allows for a higher masking ratio, which provides a greater speedup benefit while also enjoying good accuracy.

Training schedule. Our ablations thus far are based on 800-epoch pre-training. Figure 7 shows the influence of the training schedule length. The accuracy improves steadily with longer training. Indeed, we have not observed saturation of linear probing accuracy even at 1600 epochs. This behavior is unlike contrastive learning methods, e.g., MoCo v3 [9] saturates at 300 epochs for ViT-L. Note that the MAE encoder only sees 25% of patches per epoch, while in contrastive learning the encoder sees 200% (two-crop) or even more (multi-crop) patches per epoch.
4.2. Comparisons with Previous Results

Comparisons with self-supervised methods. In Table 3 we compare the fine-tuning results of self-supervised ViT models. For ViT-B, all methods perform closely. For ViT-L, the gaps among methods are bigger, suggesting that a challenge for bigger models is to reduce overfitting.

Our MAE can scale up easily and has shown steady improvement from bigger models. We obtain 86.9% accuracy using ViT-H (224 size). By fine-tuning with a 448 size, we achieve 87.8% accuracy, using only IN1K data. The previous best accuracy, among all methods using only IN1K data, is 87.1% (512 size) [67], based on advanced networks. We improve over the state-of-the-art by a nontrivial margin in the highly competitive benchmark of IN1K (no external data). Our result is based on vanilla ViT, and we expect advanced networks will perform better.

Comparing with BEiT [2], our MAE is more accurate while being simpler and faster. Our method reconstructs pixels, in contrast to BEiT that predicts tokens: BEiT reported a 1.8% degradation [2] when reconstructing pixels with ViT-B. We do not need dVAE pre-training. Moreover, our MAE is considerably faster (3.5× per epoch) than BEiT, for the reason as studied in Table 1c.

The MAE models in Table 3 are pre-trained for 1600 epochs for better accuracy (Figure 7). Even so, our total pre-training time is less than the other methods when trained on the same hardware. For example, training ViT-L on 128 TPU-v3 cores, our MAE’s training time is 31 hours for 1600 epochs and MoCo v3’s is 36 hours for 300 epochs [9].

Comparisons with supervised pre-training. In the original ViT paper [16], ViT-L degrades when trained in IN1K. Our implementation of supervised training (see A.2) works better, but accuracy saturates. See Figure 8.

Our MAE pre-training, using only IN1K, can generalize better: the gain over training from scratch is bigger for higher-capacity models. It follows a trend similar to the JFT-300M supervised pre-training in [16]. This comparison shows that our MAE can help scale up model sizes.

4.3. Partial Fine-tuning

Table 1 shows that linear probing and fine-tuning results are largely uncorrelated. Linear probing has been a popular protocol in the past few years; however, it misses the opportunity of pursuing strong but non-linear features—which is indeed a strength of deep learning. As a middle ground, we study a partial fine-tuning protocol: fine-tune the last several layers while freezing the others. This protocol was also used in early works, e.g., [65, 70, 42].

Figure 9 shows the results. Notably, fine-tuning only one Transformer block boosts the accuracy significantly from 73.5% to 81.0%. Moreover, if we fine-tune only “half” of the last block (i.e., its MLP sub-block), we can get 79.1%, much better than linear probing. This variant is essentially fine-tuning an MLP head. Fine-tuning a few blocks (e.g., 4 or 6) can achieve accuracy close to full fine-tuning.

In Figure 9 we also compare with MoCo v3 [9], a contrastive method with ViT-L results available. MoCo v3 has higher linear probing accuracy; however, all of its partial fine-tuning results are worse than MAE. The gap is 2.6% when tuning 4 blocks. While the MAE representations are less linearly separable, they are stronger non-linear features and perform well when a non-linear head is tuned.
5. Transfer Learning Experiments

We evaluate transfer learning in downstream tasks using the pre-trained models in Table 3.

Object detection and segmentation. We fine-tune Mask R-CNN [24] end-to-end on COCO [37]. The ViT backbone is adapted for use with FPN [36] (see A.3). We apply this approach for all entries in Table 4. We report box AP for object detection and mask AP for instance segmentation.

Compared to supervised pre-training, our MAE performs better under all configurations (Table 4). With the smaller ViT-B, our MAE is 2.4 points higher than supervised pre-training (50.3 vs. 47.9, APbox). More significantly, with the larger ViT-L, our MAE pre-training outperforms supervised pre-training by 4.0 points (53.3 vs. 49.3).

The pixel-based MAE is better than or on par with the token-based BEiT, while MAE is much simpler and faster. Both MAE and BEiT are better than MoCo v3 and MoCo v3 is on par with supervised pre-training.

Semantic segmentation. We experiment on ADE20K [72] using UperNet [63] (see A.4). Table 5 shows that our pre-training significantly improves results over supervised pre-training, e.g., by 3.7 points for ViT-L. Our pixel-based MAE also outperforms the token-based BEiT. These observations are consistent with those in COCO.

Classification tasks. Table 6 studies transfer learning on the iNaturalists [56] and Places [71] tasks (see A.5). On iNat, our method shows strong scaling behavior: accuracy improves considerably with bigger models. Our results surpass the previous best results by large margins. On Places, our MAE outperforms the previous best results [19, 40], which were obtained via pre-training on billions of images.

Table 4. COCO object detection and segmentation using a ViT Mask R-CNN baseline. All entries are based on our implementation. Self-supervised entries use IN1K data without labels. Mask AP follows a similar trend as box AP.

<table>
<thead>
<tr>
<th>method</th>
<th>pre-train data</th>
<th>ViT-B</th>
<th>ViT-L</th>
<th>ViT-B</th>
<th>ViT-L</th>
</tr>
</thead>
<tbody>
<tr>
<td>supervised</td>
<td>IN1K w/ labels</td>
<td>47.9</td>
<td>49.3</td>
<td>42.9</td>
<td>43.9</td>
</tr>
<tr>
<td>MoCo v3</td>
<td>IN1K</td>
<td>47.9</td>
<td>49.3</td>
<td>42.7</td>
<td>44.0</td>
</tr>
<tr>
<td>BEiT</td>
<td>IN1K+DALLE</td>
<td>49.8</td>
<td>53.3</td>
<td>44.4</td>
<td>47.1</td>
</tr>
<tr>
<td>MAE</td>
<td>IN1K</td>
<td>50.3</td>
<td>53.3</td>
<td>44.9</td>
<td>47.2</td>
</tr>
</tbody>
</table>

Table 5. ADE20K semantic segmentation (mIoU) using UperNet. BEiT results are reproduced using the official code. Other entries are based on our implementation. Self-supervised entries use IN1K data without labels.

<table>
<thead>
<tr>
<th>dataset</th>
<th>VIT-B</th>
<th>VIT-L</th>
<th>VIT-H</th>
<th>VIT-H48</th>
</tr>
</thead>
<tbody>
<tr>
<td>iNat 2017</td>
<td>70.5</td>
<td>75.7</td>
<td>79.3</td>
<td>83.4</td>
</tr>
<tr>
<td>iNat 2018</td>
<td>75.4</td>
<td>80.1</td>
<td>83.0</td>
<td>86.8</td>
</tr>
<tr>
<td>iNat 2019</td>
<td>80.5</td>
<td>83.4</td>
<td>85.7</td>
<td>88.3</td>
</tr>
<tr>
<td>Places205</td>
<td>63.9</td>
<td>65.8</td>
<td>65.9</td>
<td>66.8</td>
</tr>
<tr>
<td>Places365</td>
<td>57.9</td>
<td>59.4</td>
<td>59.8</td>
<td>60.3</td>
</tr>
</tbody>
</table>

Table 6. Transfer learning accuracy on classification datasets, using MAE pre-trained on IN1K and then fine-tuned. We provide system-level comparisons with the previous best results.

<table>
<thead>
<tr>
<th>dataset</th>
<th>IN1K</th>
<th>VIT-L</th>
<th>VIT-H</th>
<th>COCO</th>
</tr>
</thead>
<tbody>
<tr>
<td>iNat</td>
<td>48.1</td>
<td>53.6</td>
<td>58.0</td>
<td>66.0</td>
</tr>
<tr>
<td>Places365</td>
<td>44.4</td>
<td>52.8</td>
<td>50.3</td>
<td>53.2</td>
</tr>
<tr>
<td>ADE20K</td>
<td>51.8</td>
<td>53.6</td>
<td>48.0</td>
<td>51.3</td>
</tr>
</tbody>
</table>

Table 7. Pixels vs. tokens as the MAE reconstruction target. △ is the difference between using dVAE tokens and using normalized pixels. The difference is statistically insignificant.

6. Discussion and Conclusion

Simple algorithms that scale well are the core of deep learning. In NLP, simple self-supervised learning methods (e.g., [47, 14, 48, 4]) enable benefits from exponentially scaling models. In computer vision, practical pre-training paradigms are dominantly supervised (e.g. [33, 51, 25, 16]) despite progress in self-supervised learning. In this study, we observe on ImageNet and in transfer learning that an autoencoder—a simple self-supervised method similar to techniques in NLP—provides scalable benefits. Self-supervised learning in vision may now be embarking on a similar trajectory as in NLP.

On the other hand, we note that images and languages are signals of a different nature and this difference must be addressed carefully. Images are merely recorded light without a semantic decomposition into the visual analogue of words. Instead of attempting to remove objects, we remove random patches that most likely do not form a semantic segment. Likewise, our MAE reconstructs pixels, which are not semantic entities. Nevertheless, we observe (e.g., Figure 4) that our MAE infers complex, holistic reconstructions, suggesting it has learned numerous visual concepts, i.e., semantics. We hypothesize that this behavior occurs by way of a rich hidden representation inside the MAE. We hope this perspective will inspire future work.
References


