Masked Autoencoding Does Not Help Natural Language Supervision at Scale

Floris Weers Vaishaal Shankar Angelos Katharopoulos Yinfei Yang Tom Gunter
{fweers,vaishaal_shankar,a_katharopoulos,yinfeiy,tom_gunter}@apple.com

Abstract

Self supervision and natural language supervision have emerged as two exciting ways to train general purpose image encoders which excel at a variety of downstream tasks. Recent works such as M3AE [31] and SLIP [63] have suggested that these approaches can be effectively combined, but most notably their results use small (<20M examples) pre-training datasets and don't effectively reflect the large-scale regime (>100M samples) that is commonly used for these approaches. Here we investigate whether a similar approach can be effective when trained with a much larger amount of data. We find that a combination of two state of the art approaches: masked auto-encoders, MAE [37] and contrastive language image pre-training, CLIP [68] provides a benefit over CLIP when trained on a corpus of 11.3M image-text pairs, but little to no benefit (as evaluated on a suite of common vision tasks) over CLIP when trained on a large corpus of 1.4B images. Our work provides some much needed clarity into the effectiveness (or lack thereof) of self supervision for large-scale image-text training.

1. Introduction

Large scale pretraining has become a powerful tool in the arsenal of computer vision researchers to produce state of the art results across a wider variety of tasks [39,88,95,98]. However, when pre-training on tens of millions to billions of images it is difficult to rely on standard supervised methods to train models, as datasets of this size often lack reliable labels. In the presence of these massive but largely under-curated datasets, two general classes of methods to train general purpose image encoders have emerged:

1. *Self Supervised* techniques that learn visual representations from the image data alone [11,36]

2. *Natural Language Supervised* methods that utilize paired free-form text data to learn visual representations [43,69]

Due to the unique strengths and weaknesses of each approach[1], a recent flurry of work has introduced methods that combine both forms of supervision [31,56,64,78] to varying degrees of success. While each of these methods establishes some regime where the additional supervision helps, none of these "joint-supervision" methods advance state of the art in any meaningful way. Additionally, to our knowledge none of these methods have shown comparative results at the scale many large scale vision models are currently trained at (>100M examples) [43,66,69,73,80,82,98]. Furthermore, methods that use both forms of supervision start with the presumption that the additional supervision **is helpful** and either often lack clean ablations or lack evaluations in a "high accuracy" regime—leading to further confusion regarding whether a combination of these methods can actually improve the state of the art. To clarify this issue, in this work, we investigate a simple question:

> *Does a combination of self supervision and natural language supervision actually lead to higher quality visual representations?*

In order to answer this, we first introduce a straightforward baseline approach that combines standard self supervision and language supervision techniques. We combine masked auto-encoders (MAE) and contrastive language image-pretraining (CLIP) to make MAE-CLIP. We then present a careful study of the performance of MAE, M3AE, CLIP and MAE-CLIP across a wide variety of tasks in two distinct regimes: a "low-sample" [2] 11.3 million example regime and a "high-sample" 1.4 billion example regime. We train self-supervised and language-supervised methods using the same pre-training datasets under the assumption that we have no knowledge about downstream tasks. Our experiments show:

1. In the low sample size regime, without changing the final pooling operation in the network, we observe a large performance improvement, namely 6% on ImageNet [18] and 4% on VTAB [105]. However, when

[1] Self supervised methods can learn representations without labels, but natural language supervision learns better representations. Natural language supervised methods rely on quality of captions

[2] We note that what low sample means has changed substantially over the last few years

we modify the pooling operation, the improvement **substantially** decreases to around 1% on both ImageNet and VTAB.

2. In the high sample size regime, there is virtually no difference in performance between MAE-CLIP and CLIP across ImageNet, VTAB, and VQA tasks.

We believe our work is the first careful study of this form and contextualizes recent progress in both self-supervision and natural language supervision.

The rest of the paper is organized as follows: In Section 2, we cover related work in the areas of self supervision and natural language supervision. In Section 3, we give an overview of the baseline methods we study, MAE, M3AE, CLIP and our new baseline MAE-CLIP. Then we present and analyse our small scale and large scale experimental findings in Sections 4 and 5. Finally, we discuss potential explanations for our findings and some future work in 6.

2. Related Work

Our work combines **natural language supervision** and **self-supervision** in a **multi-task** approach to visual encoding, and so research from these three areas is relevant.

Natural language supervision for visual encoding covers a variety of approaches that assume access to datasets of images or videos associated with text. Some of the most successful use a contrastive pairwise alignment signal applied to very large batch and dataset sizes [68]. This large batch size means hard negative pairs can be produced by random sampling, avoiding the need for a memory bank [12, 13, 38] or momentum distillation [15]. FILIP [96] further improves zero-shot performance for some tasks by using a more fine-grained elementwise contrastive loss. Image captioning has also shown promise as a pre-training task [20, 90], with [20] in particular demonstrating strong data efficiency. UNITER [16] pursues a similar idea, applying a generative loss to both image and text modalities. [2, 16, 23, 41, 47, 51, 61, 81, 84, 97] apply the masked patch prediction problem from [21, 37, 50] to a joint image-text data space. These approaches either use a pre-trained convolutional neural network (CNN) to generate region of interest (RoI) proposal encodings, or predict the labelled class of the masked patch instead of the raw pixel values (or the quantized patch ID). UNIMO-2 [53] attempts to ground image and text patches for a single data example into a top-k quantized set of embeddings, before passing the grounded embeddings along with the raw non-grounded embeddings into a decoder, targeting a loss based on masked language modeling (MLM) and image-text-matching (ITM).

Self-supervised approaches for visual encoding may be loosely categorized into those that target consistency constraints between multiple views of the same scene and those that attempt visual reconstruction on corrupted or conditioning representations of images. Consistency constraints are often derived through data-augmentation applied to a single real image [6, 10, 14, 35, 102], but these may also take the form of e.g. a temporal constraint if video data is available [89], or be applied at an image-patch level by making a smoothness assumption and selecting neighboring patches as 'consistent' whilst treating distant patches (or those sourced from another image) as negative examples [101]. Forms of denoising autoencoder (DAE) [87] are popular as a means of self-supervision, and have been investigated at a variety of scales; iGPT [9] learns to generate images as a row-major flattened sequence of pixels, whilst [65], [72], and [26] compress images into a short sequence of discrete codes before then regenerating and scoring at a pixel level. BEiT [3] first encodes an image into a sequence of discrete codes before masking and then predicting a subset of the codes. MAE [37] drops the discretization, instead predicting raw pixel loss for a subset of the encoded patches in a manner strongly reminiscent of BART [50]. SplitMask [25] applies both a masked image patch prediction loss and a pairwise contrastive loss, using the masked example as well as the inverted masked example to form each positive pair. Meanwhile, M3AE [24] and VL-BEiT [4] both propose a masked patch prediction problem applied to image and text modalities jointly. Recent concurrent work shows that while masking does *not* help generalization, the masking can be used to speed up training by dropping the masked tokens during the forward and backward pass [55]. EVA [29] shows that doing masked image modeling to predict the output *embeddings* (rather than pixels) of the masked image can be an effective pretraining task when combined with natural language supervision. We note for EVA the masked image modeling occurs only on smaller well curated datasets such as ImageNet, Conceptual Captions and COCO.

Multi-task methods for pre-training visual encoders are a highly active area of research. CoCA [99] combines the cross-modality contrastive task from [68] with image captioning in the style of [20], and by pre-training at very large scale show that the resulting model is more performant than prior art across a very broad array of downstream visual understanding tasks. SLIP [64] combines a SimCLR style self supervision loss with the standard CLIP contrastive loss to train a jointy self-supervised and natural language supervised model, however while they show large performance gains they are all in a low accuracy regime (below 40% Top-1 accuracy on ImageNet). Florence [100] employs a diverse array of tasks and datasets at pre-training time (including object detection, CLIP-style contrastive modeling, Visual Question Answering (VQA), and supervised classification). ALBEF [52] builds on a CLIP-style architecture by incorporating a single-stream encoder that consumes both

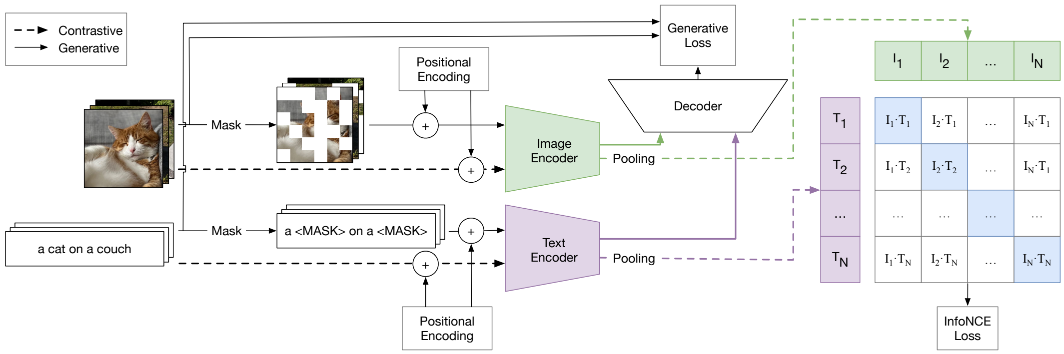

Figure 1. **MAE-CLIP**: CLIP [68] augmented with a generative decoder in the style of MAE [31].

modalities. It uses a masked language modeling loss for the text modality and an image-text matching loss for within-batch hard-negatives, mined according to contrastive similarity. FLAVA [79] also uses masked-image modeling, and uses both single and joint modality encoders, with additional single-modality datasets (i.e. not just image-text pairs) that allow it to generalize to longer text inputs and visual data which is unlikely to be captioned. X-VLM [103] utilizes image data which includes object-labelled bounding boxes to extend the image-text matching and contrastive losses to cover within-image regions as well as whole-image captions. They also include a masked language modelling objective, achieving impressive performance on many zero-shot downstream tasks—although the requirement for object bounding box data may not scale well for large pre-training datasets. SupMAE [57] adds a supervised labeling task to the MAE architecture, demonstrating that this results in better data efficiency. MVP [91], meanwhile, finetunes a pre-trained CLIP backbone using MIM, and demonstrates that starting with a pre-trained model improves downstream visual recognition performance. **MAE-CLIP** is reminiscent of M3AE (with the addition of a contrastive task), ALBEF (but applying the masked prediction task to both modalities, dropping the momentum distillation), and most of all FLAVA (without the single-modality masked-patch-prediction tasks, and with the use of high ratio random masking following MAE and M3AE).

3. Background

3.1. Contrastive language-image pre-training

CLIP [68] and ALIGN [44] both demonstrated that contrastive image-language supervision applied at scale was capable of producing an image encoder that excels at a range of downstream tasks—often in a zero-shot setting. In our implementation, we begin by following the design outlined in [68]. Specifically, we use per-modality encoders to produce a single $\ell 2$-normalized dense embedding vector for each image or text input, before applying a pairwise InfoNCE [65] loss to a large global batch of paired image-text embeddings, using all non-paired examples as hard negatives. We define the image-to-text loss L_{i2t} and the text-to-image loss L_{t2i} as:

$$L_{i2t} = -\frac{1}{N}\sum_{j}^{N}\log\frac{\exp(x_j^T y_j/\sigma)}{\sum_{k=1}^{N}\exp(x_j^T y_k/\sigma)} \qquad (1)$$

$$L_{t2i} = -\frac{1}{N}\sum_{j}^{N}\log\frac{\exp(y_j^T x_j/\sigma)}{\sum_{k=1}^{N}\exp(y_j^T x_k/\sigma)}, \qquad (2)$$

where N is the global batch size, x_j is the normalized embedding of the image for the j-th pair and y_k is the normalized embedding of the text in the k-th pair in the batch. σ is a learnable temperature parameter. The total contrastive loss $L_c = \frac{1}{2}(L_{t2i} + L_{i2t})$ is the average of these two losses.

We use transformer encoders for both modalities, however, differently to [68] we explore several strategies for aggregating the image-patch or text-token outputs into a single embedding (see Section 4.4). We also eschew the autoregressive masking strategy for the text encoder, instead allowing full bi-directional self attention following [44].

3.1.1 Pooling

In CLIP [68] the authors use a separate 'CLS' token projected through the network as the overall image representation. On the other hand, ALIGN [44] use global-average pooling over the encoded visual features. Whilst both approaches produced good downstream evaluation results, recent work [8, 71, 104], suggested that the choice of pooling strategy can strongly influence the quality of visual semantic embeddings. In particular, [71] demonstrated that this effect is present for CLIP-like visual encoders. Noting this, we opt to investigate three pooling strategies: 1. the default multihead-attention pooling (MAP), 2. global average pooling (GAP)—both described in [104]—as well as 3. non-maximal suppression pooling (MAX) as in [8, 71].

3.2. Masked Autoencoders

In MAE [37] the authors demonstrate a simple technique for self-supervised image-encoder pre-training that—to our knowledge—is still considered state-of-the-art. They use a

ViT [22] encoder-decoder architecture and apply it to heavily masked input images. The input to the encoder consists of the visible, unmasked image patches, first embedded via a linear projection before additive positional embeddings are applied, and then the result is fed through the encoder's transformer layers. They demonstrate that a very high patch masking ratio is critical to achieving good performance, and usually retain only 25% of the image. The decoder consumes the output of the encoder, as well as a learned 'masked-patch' embedding, which is included to represent each masked token. Positional embeddings are also reapplied, to ensure that spatial information is passed through the decoder for the masked patches. The output of the decoder at the masked positions is then measured against ground truth using a simple mean-squared-error loss. As in [37] we experiment with predicting both normalized and un-normalized patch values, finding that predicting the normalized patch value slightly improves the performance of our MAE implementation.

Masking across modalities: BART [50] applies a similar strategy to text data, but uses a cross-entropy loss over the masked-token output distribution and a far lower masking ratio (typically around 15%). M3AE [31] extends MAE and BART to incorporate inputs from both text and image modalities, using per-modality input and output projections. Otherwise the encoder-decoder architecture resembles MAE, with the addition of learned modality-indicating embeddings to each transformer input.

The M3AE loss relies only on the contents of the image-text pairs, and comprises mean-squared error for the masked image patches (L_{gen_i}) and cross-entropy over the vocabulary for the masked text tokens (L_{gen_t}):

$$L_{\text{gen\_i}} = \frac{1}{N}\sum_{j}^{N}(p_j - P_j)^2 \tag{3}$$

$$L_{\text{gen\_t}} = -\frac{1}{N}\sum_{j}^{N}\log\frac{\exp(t_{n,T_n})}{\sum_{c=1}^{C}\exp(t_{n,c})}, \tag{4}$$

where p_j and P_j refer to the predicted and ground truth pixel value respectively and t_n and T_n refer to the predicted token distribution and the ground truth token, with C as the number of unique tokens. For the image reconstruction loss, we normalize the ground-truth per-patch, following [37].

3.3. MAE-CLIP

As the name suggests, MAE-CLIP attempts to incorporate aspects of MAE/M3AE into CLIP through the addition of a single dual-modality transformer decoder (as in M3AE) which consumes the output of the CLIP encoders. We show this in Figure 1.

3.3.1 Components

The model architecture is similar to ALBEF [52] and FLAVA [78], consisting of three components.

An image encoder: Following ViT [22], we divide the input image into equally-sized, non-overlapping patches. After applying a linear projection and adding a 2-D position encoding, we feed the per-patch representations through a number of transformer layers [86].

A text encoder: Following ALIGN [43], our text encoder is based on BERT [21], in which text is first tokenized and embedded, and a 1-D trainable position encoding is added. Differently to BERT, we use pre-layernorm [94] and initialize the parameters according to the simplified scheme outlined in [70].

A cross-modality decoder: The decoder receives per-element encoded image and text representations from the encoders for both masked and un-masked elements. For the image modality, masked patches are added by replacing their values with a shared trainable mask token. Positional encodings are once again added to all elements, following [37]. We also add a per-modality trainable encoding to allow the decoder to easily distinguish between the two modalities. The decoder uses the same transformer implementation as the encoders. The output of the final decoder layer is then projected into a per-patch-modality output space and the overall loss is computed.

3.3.2 Losses

MAE-CLIP employs the losses of both CLIP (Eq. 1, 2) and M3AE (Eq. 3, 4). Our final loss is a weighted sum of the losses from each task, as follows:

$$L = \frac{1}{2}(L_{t2i} + L_{i2t}) + w_i \cdot L_{\text{gen\_i}} + w_t \cdot L_{\text{gen\_t}} \tag{5}$$

where w_i and w_t are scalars used to control the relative weight of the generative losses. We always provide paired image-text inputs to the model, leaving it to future work to explore the benefits of also incorporating single modality inputs and reconstruction losses, at scale.

To avoid severely impacting the CLIP loss via computing it with masked inputs, we run the encoders twice: once with full unmasked input in order to compute the CLIP loss, before then making a second pass using only the unmasked input to compute the M3AE loss. Finally, we compute the weighted sum of the two losses, and use this to calculate the overall update. [3]

[3] Alternative, more memory efficient, strategies exist for computing the overall updates (e.g. calculate per-loss gradients, or perform round-robin updates), but prior work [1] has demonstrated that some of these techniques may negatively impact the overall training result.

4. Experiments

In this section, we present our experimental results to study how the addition of self-supervision affects the representation quality of natural language supervised models. We study standard masked auto-encoders (MAE), multimoddal masked auto-encoders (M3AE), contrastive language image pretraining (CLIP) and our newly introduced baseline method (MAE-CLIP), which combines mask based self-supervision with contrastive language-image learning.

4.1. Experimental Setup

Architecture: Experiments are performed using an image encoder based on the architecture described as ViT-B-16 in [93], in combination with a 12-layer, 512-wide text encoder with 8 attention heads. For the decoder, we use another 12-layer, 512-wide transformer with 8 attention heads, identical to the text encoder. We also use a byte-pair encoding tokenizer, fit to the OpenWebText corpus [32]. When training MAE-CLIP, we replace the random masking by a similarity masking strategy, which makes use of the element-wise CLIP similarity scores to select the masked element for the reconstruction task. [4]

Pretraining Datasets: Our analysis is divided into two sections. First, in sections 4.2 to 4.4, we present our study on the "low-sample" regime using the combination of CC12M [7] and CC3M [77] excluding all images whose text contains a *[PERSON]* tag[5]. The final dataset contains 11.3 million comparatively high quality image-text pairs. We refer to this dataset as *CC* or "small" scale.

Pretraining configuration: We use a minibatch size of $16,384$ samples and train using the AdamW optimizer [60] with decoupled weight decay. We use a base learning rate of 5×10^{-4} and train using cosine learning rate decay [59] with an initial linear warmup of 200 steps. We reduce the β_2 decay factor to 0.98, and increase ϵ to 10^{-6}, finding that these changes improved rate of convergence for all models trained. To increase training convergence speed, we switch from a local (per-GPU) to a global contrastive loss after 500 steps. During the global contrastive loss phase, we set the generative text and image loss weights to $w_t = 0.05$ and $w_i = 0.1$ respectively. For all models which rely on masking, we use a 75% masking ratio, consistent with [31, 37], as we did not find an alternative that improved downstream results. In total, we train our models on *CC* for 32 epochs, which corresponds to $\sim 22,000$ steps. For our M3AE training runs, we use per-modality encoders so as to produce a baseline that is flop and parameter matched to our MAE-CLIP runs.

[4]We compare the similarity masking strategy with the original random masking strategy on various tasks, the results show they perform similarly.

[5]Preliminary experiments on CLIP suggest removing such data leads to better performance.

Models	Zero-shot	Linear Probing
MAE	–	33.9
M3AE	–	52.5
CLIP	29.7	52.6
MAE-CLIP	33.8	58.9

Table 1. ImageNet classification with zero-shot transfer or linear probing after pretraining on the *CC* dataset (11.3M images). MAE-CLIP significantly improves the classification performance of CLIP in the small scale regime.

4.2. Image Classification

We evaluate the quality of the representations learned from pre-training on *CC* by measuring the image classification performance, either by zero-shot transfer, or by training a linear classifier using the predicted visual features. For zero-shot classification, we use the average of 80 prompts and follow the routine prescribed in [69]. To train the linear classifier, we pre-compute features using the visual encoder[6] and run AdamW for 20 to 80 epochs with a learning rate of 0.01.

Table 1 shows the zero-shot and linear probing classification results on the ImageNet [19] dataset while Table 2 shows our linear probing results on most (17) of the VTAB [106] benchmark datasets[7]. Initially, we observe that the combination of self-supervision and natural language supervision provides a consistent and substantial improvement over either form of supervision by itself. This result concurs with previous works such as [31] and [64], which show that self-supervision aids natural language supervision.

4.3. VQA

Subsequently, we compare the performance of MAE-CLIP to the rest of the baselines on the visual question answering task using three datasets. CLEVR [45], VQAv2 [34] and GQA [42]. VQA assesses the model's multimodal reasoning capabilities as well as its visual grounding. In order to finetune our models for VQA, we freeze the image and text encoders and either randomly reinitialize the decoder in models such as MAE-CLIP or M3AE or add a new identical decoder for CLIP. During finetuning, we use a layer-wise learning rate decay, following [3, 17, 37]. For CLEVR and VQAv2 we finetune the decoder for 50 epochs, while for GQA we finetune for 5 epochs on the "train-all" split and 2 epochs on the "train-balanced" split following the protocol of [54]. In all cases, we treat the problem as a classification problem of selecting one answer out of the set of possible answers for each dataset.

In Table 3 we observe that self-supervision combined

[6]We use no data augmentation here

[7]We do not include results for either the diabetic retinopathy or the Sun397 tasks, due to licensing issues. More details in the supplementary material

Table 2. Linear probing accuracy (%) on classification tasks. All models are trained on the *CC* Dataset (11.3M images). (● VTAB/natural, ● VTAB/specialized and ● VTAB/structured.)

	Caltech101 ●	CIFAR-100 ●	DTD ●	Flowers102 ●	Pets ●	SVHN ●	EuroSAT ●	Camelyon ●	Resisc45 ●	Clevr/Closest ●	Clevr/Count ●	DMLab ●	dSprites/Ori ●	dSprites/Loc ●	KITTI/Dist ●	sNORB/Azim ●	sNORB/Elev ●	Average
MAE	75.3	56.2	58.5	70.2	37.4	71.9	96.2	82.8	84.0	65.1	60.6	42.8	37.9	79.9	36.7	37.4	64.5	62.2
M3AE	85.6	66.6	69.4	86.5	58.4	69.2	97.4	82.9	90.7	69.4	71.2	48.2	49.8	83.3	42.6	34.3	73.4	69.3
CLIP	84.2	62.8	57.7	81.6	69.7	52.5	95.5	82.6	86.7	53.7	52.9	44.8	45.9	61.3	45.7	31.0	42.6	61.8
MAE-CLIP	89.8	66.2	64.8	86.5	74.5	56.7	95.4	81.6	86.9	53.6	59.6	45.9	44.6	68.6	48.0	36.1	51.0	65.3

Model	CLEVR	VQAv2	GQA
MAE	93.5	45.7	46.1
M3AE	**97.5**	55.8	50.9
CLIP	87.5	55.6	50.0
MAE-CLIP	96.0	**58.5**	**52.2**

Table 3. VQA finetuning results for models pre-trained on CC Dataset (11.3M images). We train a new decoder for all methods while keeping the encoders frozen. MAE-CLIP performs significantly better than either only self-supervised or language supervised methods by themselves.

Model	Pooling	ImageNet ZS	ImageNet LP	VTAB LP	VQA FT
CLIP	MAP	29.7	52.6	61.8	64.4
MAE-CLIP	MAP	33.8	58.9	65.3	68.9
CLIP	GAP	29.3	59.8	70.6	65.1
MAE-CLIP	GAP	33.5	62.5	**71.7**	**69.2**
CLIP	MAX	33.1	62.3	68.7	63.5
MAE-CLIP	MAX	**35.2**	**63.2**	69.1	68.5

Table 4. Results of CLIP and MAE-CLIP with different pooling options trained on CC dataset (11.3M images). We compare the models on ImageNet Zero-Shot (ZS), ImageNet Linear-Probing (LP), VTAB 17-task Linear-Probing average, and VQA 3 tasks Fine-Tuning (FT) average.

with natural language supervision (MAE-CLIP) performs consistently better than either MAE or CLIP by themselves. Combining information from both modalities is evidently critical for effectively answering visual questions as the performance of MAE on VQAv2 and GQA is several percentage points lower than the rest of the methods. Interestingly, methods that employ a form of self-supervision perform significantly better on CLEVR. We argue that this discrepancy is due to CLEVR requiring object localization and spatial reasoning which benefits from this form of self-supervision.

4.4. Pooling Analysis

To better understand whether the gap between MAE-CLIP and CLIP we observed in the previous section is fundamental, we investigate the effect of the pooling operator on different downstream tasks. Table 4 compares the performance of CLIP and MAE-CLIP with the standard multi-head attention pooling (MAP), global average pooling (GAP) and max pooling (MAX) on all image classification tasks and visual question answering. We note that GAP and MAX pooling perform substantially better than the standard MAP pooling across ImageNet, VTAB and VQA tasks. A hypothesis posed by [71] is that the alternative pooling operators lead to better *perceptual grouping* in the visual representation. We study this with a qualitative analysis in Section 6.1.

5. Experiments at Scale

At small-scale, we showed that self-supervision combined with natural language supervision marginally improves the quality of the learned visual representations, and that the choice of image-encoder pooling strategy can heavily influence results. We now investigate whether these conclusions still hold when training on a much larger, 1.4B example dataset.

5.1. Experimental Setup

We follow the configuration described in Section 4.1, with a few key differences, noted below.

Pretraining Dataset: We combine a 2.2B example web-crawled image-text pair dataset, termed the English-Web-Image-Text dataset (EWIT-2.2B), with LAION-400M [74], CC3M [77], CC12M [7] and an internal high-quality image-text pair dataset containing approximately 134M image text pairs which we term the High Quality Image Text Pairs Dataset (HQITP-134M). We globally deduplicate this—keyed by image bytes—reducing the number of image-text pairs significantly, to yield a final 1.4B examples. Details provided in the supplementary material. In the rest of the paper we refer to this as the *web-crawled* or large-scale dataset.

Pretraining configuration: We increase the learning rate warm-up to $1,000$ steps, and compute a local contrastive loss for the first $10,000$ steps, as during early training this improves the models rate of convergence. We train for a total of $480,000$ steps, which corresponds to 6 full passes through the dataset. We do not increase the batch

Table 5. Linear probing accuracy (%) on classification tasks. Models are all trained on our *web-crawled* dataset (1.4B images). (● VTAB/natural, ● VTAB/specialized and ● VTAB/structured.) In the large scale pretraining regime, the difference between MAE-CLIP and CLIP is reduced to $< 1\%$. * At evaluation time, our M3AE model was 50% trained, so that performance may improve further.

	Caltech101 ●	CIFAR-100 ●	DTD ●	Flowers102 ●	Pets ●	SVHN ●	EuroSAT ●	Camelyon ●	Resisc45 ●	Clevr/Closest ●	Clevr/Count ●	DMLab ●	dSprites/Ori ●	dSprites/Loc ●	KITTI/Dist ●	sNORB/Azim ●	sNORB/Elev ●	Average
M3AE*	93.0	74.8	78.2	95.4	81.2	69.6	97.3	84.8	92.7	65.7	74.7	51.0	52.7	80.7	47.5	36.2	68.8	73.2
CLIP	94.9	78.4	80.0	97.3	86.9	59.0	94.1	82.3	92.7	45.6	62.1	46.0	46.1	53.3	50.9	20.3	35.8	66.2
CLIP_{MAX}	96.1	81.0	80.9	97.3	89.9	65.7	96.0	83.2	94.1	52.8	67.8	49.9	59.5	67.6	41.2	23.4	45.8	70.1
$\text{MAE-CLIP}_{\text{MAX}}$	95.8	79.2	81.5	96.8	88.2	62.1	95.8	81.8	93.0	52.0	66.9	49.6	53.7	72.5	53.0	32.3	45.4	70.6
CLIP_{GAP}	95.8	80.5	81.6	97.6	88.7	66.0	97.0	84.4	93.3	56.7	71.4	53.3	58.0	70.1	50.6	38.3	55.1	72.9
$\text{MAE-CLIP}_{\text{GAP}}$	95.4	79.3	82.2	97.4	88.6	72.8	96.6	84.5	93.5	57.5	73.6	52.7	57.5	71.2	51.6	45.6	55.2	73.8

Models	**Zero-shot**	**Linear Probing**
M3AE*	–	69.3
CLIP_{GAP}	61.8	75.9
CLIP_{MAX}	63.7	77.5
$\text{MAE-CLIP}_{\text{GAP}}$	57.4	75.7
$\text{MAE-CLIP}_{\text{MAX}}$	60.9	76.6

Table 6. ImageNet classification after pretraining on *web-crawled* dataset (1.4B images). In the large scale regime, self-supervision does not complement natural language supervision.

Model	**CLEVR**	**VQAv2**	**GQA**
M3AE	**96.9**	59.9	53.3
CLIP_{GAP}	87.8	**61.8**	55.0
CLIP_{MAX}	89.5	60.6	53.6
$\text{MAE-CLIP}_{\text{GAP}}$	92.8	61.9	**55.3**
$\text{MAE-CLIP}_{\text{MAX}}$	93.9	61.5	53.7

Table 7. VQA finetuning results on *web-crawled* dataset (1.4B images). Similar to Table 15a, for large pretraining datasets, self-supervision is not improving the downstream VQA performance.

size, and use the same learning rate schedule as for *CC*.

Model variants: We train M3AE as well as both GAP and MAX variants of CLIP and MAE-CLIP, having noted in the small-scale regime that these are the highest performing architectures and pre-training strategies. We do not train MAE at scale due to resource limitations, and because its performance was significantly lower than the models that utilize natural language supervision—a finding that is consistent with prior literature [31].

5.2. Image Classification

Table 15a provides encoder performance for both zero-shot and linear-probing on ImageNet [19]. We note that at scale, MAE-CLIP typically shows worse performance than CLIP across both pooling strategies and tasks. We hypothesize that this may be due to the fact that masked-patch-prediction results in model capacity being lost to the task of predicting semantically irrelevant patches, meaning that we lose some of the benefits provided by the larger scale dataset (despite tuning the relative loss weights). If we are correct in this hypothesis then it is possible that a larger model capacity would help to alleviate this issue. We leave this to a future investigation. Meanwhile, in Table 14 we once again explore the encoder performance through linear-probing for VTAB [106]. We note that GAP provides a consistent improvement over MAX across both CLIP and MAE-CLIP models, and that M3AE is also a strong performer—as was the case in the small-scale regime. M3AE provides particular benefit to the "structured" VTAB tasks, however underperforms (sometimes by a large margin) on the "natural" tasks. MAE-CLIP ends up in between the two—doing better than CLIP but worse than M3AE on the structured tasks, and worse than CLIP but better than M3AE on the natural tasks.

5.3. VQA

We follow the same finetuning strategy as before, with results in Table 15b. Notably, we do not see a consistent difference across tasks for CLIP vs MAE-CLIP, with the possible exception of CLEVR [45], where it is clear that M3AE is by far the best performing approach—and its benefits improve the MAE-CLIP results on that task. GAP is perhaps a marginally stronger pooling strategy as measured against MAX, however the difference is smaller than it was in the small-scale regime.

6. Analysis and Discussion

While our results present a pessimistic view on combining self-supervision with natural language supervision, we don't have a clear reason as to *why* self-supervision fails at scale. Here we present two hypotheses that may be explored in future work.

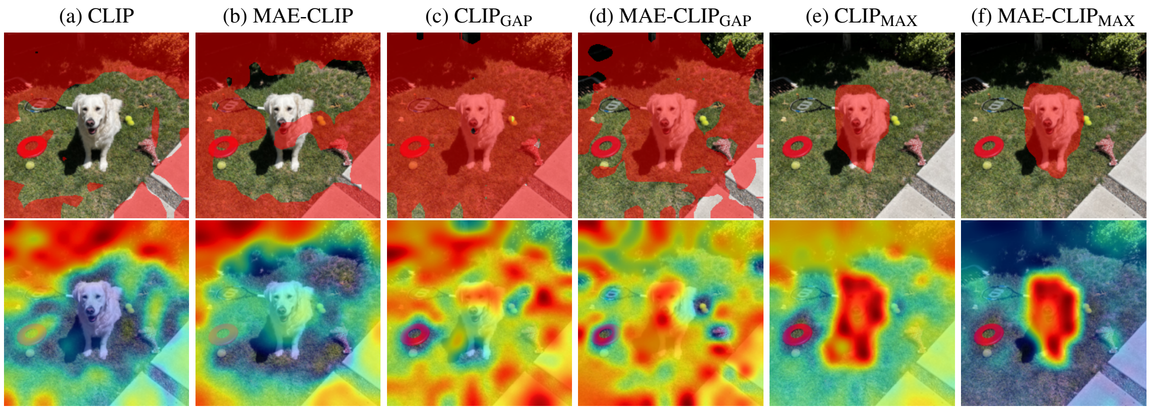

Figure 2. Zero-shot segmentation masks (top row) and GradCAM [75] visualizations (bottom row) for the prompt "a photo of a dog" for CLIP and MAE-CLIP. Self-supervision qualitatively improves the visual grounding of the model, however, the choice of pooling operator has the largest effect, with max pooling producing significantly better results [71].

6.1. Visual Grounding

Visual grounding measures how well a particular representation or network can *localize* objects within an image. It is thought that representations with better visual grounding will excel at more general purpose tasks. In section 4, we found that pooling operator had a larger effect on performance than the addition of MAE. We qualitatively study the effect of both pooling and MAE on the visual grounding of the CLIP image and text encoders. We extract zero-shot segmentation masks and GradCAM [75] visualizations for the prompt "a photo of a dog", shown in Figure 2. Details are provided in the supplementary material. We observe that, with respect to visual grounding, the pooling operator plays a much more significant role than self-supervision [71]. In particular, CLIP$_{MAX}$ correctly segments the dog in the picture while CLIP focuses only in the background. However, both for CLIP and CLIP$_{MAX}$ we observe only an incremental improvement of the localization of the dog in the image when self-supervision is added. Unfortunately, even though the GAP variants of MAE-CLIP and CLIP achieve the best performance on VTAB, they do not also exhibit the most visual grounding. We hope that a more thorough future analysis can elucidate the relationship between self supervision and visual grounding.

6.2. Dataset Diversity

Alternatively, it may be that self-supervision and natural language supervision excel for entirely different parts of the dataset diversity-size spectrum. Historically, the strongest self-supervised visual-encoder baselines (when compared to supervised methods) include methods such as: SimCLR [11], SimCLRV2 [12] and masked auto encoders [36], which all *excel* on ImageNet—a somewhat diverse but a relatively "clean" and object-centric dataset with 1000 disjoint classes. On less diverse datasets such as Cifar-10 [49], specialized self supervised methods such as ReMixMatch [5] have shown excellent performance in the extremely low data regime. Outside of vision, MAEs have shown near state-of-the-art performance on datasets such as Audioset [30,40], a clean 2M example, 632 class audio/video dataset. Recent works which perform controlled experiments scaling self supervised methods to massive datasets [62] show that self-supervised methods achieve numbers roughly 10% worse (as measured by ImageNet linear probe) as compared to natural language supervised methods at similar data scale.

On the other hand, natural language supervised models seem to only show competitive performance on massive and diverse datasets, CLIP [69] was trained on 400M examples, ALIGN [43] was trained on 1.8B examples and BASIC [66] was trained on 6.6B examples. Recently, [28] performs controlled natural language supervision at a smaller scale, showing that natural language supervised models are 5-10% worse (on ImageNet) than their supervised counterparts.

While all of these past results only pose circumstantial and not rigorous experimental evidence, we believe it poses a fruitful line of future work on studying the scaling trends of self supervised methods and contrasting them with other forms of supervision.

7. Acknowledgement

We are grateful to many colleagues for helpful advice and guidance throughout this project. We especially thank Jonathan Shlens, Brandon McKinzie, and Jason Ramuparam for feedback on paper drafts, Ruoming Pang for useful comments and suggestions early in the project, Tom Nickson for his training infrastructure support and Albin Madappally Jose for his assistance in building the training dataset. We also extend thanks to Dr. Nicola Strisciuglio for his clarifying input on earlier incarnations of this work, and Dr. Samuel Albanie for several engaging discussions and helpful pointers to related literature.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a Visual Language Model for Few-Shot Learning. Technical Report arXiv:2204.14198, arXiv, Apr. 2022. arXiv:2204.14198 [cs] type: article. 4

[2] Tarik Arici, Mehmet Saygin Seyfioglu, Tal Neiman, Yi Xu, Son Train, Trishul Chilimbi, Belinda Zeng, and Ismail Tutar. MLIM: Vision-and-Language Model Pre-training with Masked Language and Image Modeling. *arXiv:2109.12178 [cs]*, Sept. 2021. arXiv: 2109.12178. 2

[3] Hangbo Bao, Li Dong, and Furu Wei. BEiT: BERT Pre-Training of Image Transformers. *arXiv:2106.08254 [cs]*, June 2021. arXiv: 2106.08254. 2, 5

[4] Hangbo Bao, Wenhui Wang, Li Dong, and Furu Wei. VL-BEiT: Generative Vision-Language Pretraining. Technical Report arXiv:2206.01127, arXiv, June 2022. arXiv:2206.01127 [cs] type: article. 2

[5] David Berthelot, Nicholas Carlini, Ekin D. Cubuk, Alex Kurakin, Kihyuk Sohn, Han Zhang, and Colin Raffel. Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring. *CoRR*, abs/1911.09785, 2019. 8

[6] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. DINO: Emerging Properties in Self-Supervised Vision Transformers. *arXiv:2104.14294 [cs]*, May 2021. arXiv: 2104.14294. 2, 16

[7] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12M: Pushing Web-Scale Image-Text Pre-Training To Recognize Long-Tail Visual Concepts. Technical Report arXiv:2102.08981, arXiv, Mar. 2021. arXiv:2102.08981 [cs] type: article. 5, 6, 14

[8] Jiacheng Chen, Hexiang Hu, Hao Wu, Yuning Jiang, and Changhu Wang. Learning the best pooling strategy for visual semantic embedding. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3

[9] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. iGPT: Generative Pretraining From Pixels. In *Proceedings of the 37th International Conference on Machine Learning*, pages 1691–1703. PMLR, Nov. 2020. ISSN: 2640-3498. 2

[10] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. SimCLR: A Simple Framework for Contrastive Learning of Visual Representations. In *Proceedings of the 37th International Conference on Machine Learning*, pages 1597–1607. PMLR, Nov. 2020. ISSN: 2640-3498. 2

[11] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. *CoRR*, abs/2002.05709, 2020. 1, 8

[12] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. SimCLRv2: Big Self-Supervised Models are Strong Semi-Supervised Learners. *arXiv:2006.10029 [cs, stat]*, Oct. 2020. arXiv: 2006.10029. 2, 8

[13] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. MoCov2: Improved Baselines with Momentum Contrastive Learning. *arXiv:2003.04297 [cs]*, Mar. 2020. arXiv: 2003.04297. 2

[14] Xinlei Chen and Kaiming He. SimSiam: Exploring Simple Siamese Representation Learning. *arXiv:2011.10566 [cs]*, Nov. 2020. arXiv: 2011.10566. 2

[15] Xinlei Chen, Saining Xie, and Kaiming He. MoCov3: An Empirical Study of Training Self-Supervised Vision Transformers. *arXiv:2104.02057 [cs]*, Aug. 2021. arXiv: 2104.02057. 2

[16] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. UNITER: UNiversal Image-TExt Representation Learning. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, Lecture Notes in Computer Science, pages 104–120, Cham, Sept. 2020. Springer International Publishing. 2

[17] Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. Technical Report arXiv:2003.10555, arXiv, Mar. 2020. arXiv:2003.10555 [cs] type: article. 5

[18] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1

[19] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, June 2009. ISSN: 1063-6919. 5, 7

[20] Karan Desai and Justin Johnson. VirTex: Learning Visual Representations from Textual Annotations. *arXiv:2006.06666 [cs]*, Sept. 2021. arXiv: 2006.06666. 2

[21] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]*, May 2019. arXiv: 1810.04805. 2, 4

[22] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. ViT: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv:2010.11929 [cs]*, June 2021. arXiv: 2010.11929. 4

[23] Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Pengchuan Zhang, Lu Yuan, Nanyun Peng, Zicheng Liu, and Michael Zeng. METER: An Empirical Study of Training End-to-End Vision-and-Language Transformers. *arXiv:2111.02387 [cs]*, Nov. 2021. arXiv: 2111.02387. 2

[24] Jiali Duan, Liqun Chen, Son Tran, Jinyu Yang, Yi Xu, Belinda Zeng, and Trishul Chilimbi. Multi-modal Alignment using Representation Codebook. *arXiv:2203.00048 [cs]*, Mar. 2022. arXiv: 2203.00048. 2

[25] Alaaeldin El-Nouby, Gautier Izacard, Hugo Touvron, Ivan Laptev, Hervé Jegou, and Edouard Grave. SplitMask: Are Large-scale Datasets Necessary for Self-Supervised Pre-training? *arXiv:2112.10740 [cs]*, Dec. 2021. arXiv: 2112.10740. 2

[26] Patrick Esser, Robin Rombach, and Björn Ommer. VQ-GAN: Taming Transformers for High-Resolution Image Synthesis. *arXiv:2012.09841 [cs]*, June 2021. arXiv: 2012.09841. 2

[27] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The Pascal Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision*, 88(2):303–338, June 2010. 19

[28] Alex Fang, Gabriel Ilharco, Mitchell Wortsman, Yuhao Wan, Vaishaal Shankar, Achal Dave, and Ludwig Schmidt. Data determines distributional robustness in contrastive language image pre-training (clip), 2022. 8

[29] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale, 2022. 2

[30] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 776–780. IEEE, 2017. 8

[31] Xinyang Geng, Hao Liu, Lisa Lee, Dale Schuurmans, Sergey Levine, and Pieter Abbeel. M3AE: Multimodal Masked Autoencoders Learn Transferable Representations. Technical Report arXiv:2205.14204, arXiv, May 2022. arXiv:2205.14204 [cs] type: article. 1, 3, 4, 5, 7, 14, 21

[32] Aaron Gokaslan and Vanya Cohen. Openwebtext corpus. `http://Skylion007.github.io/OpenWebTextCorpus`, 2019. 5

[33] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, Large Minibatch SGD: Training ImageNet in 1 Hour. *arXiv:1706.02677 [cs]*, Apr. 2018. arXiv: 1706.02677. 15

[34] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. VQA2: Making the v in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. pages 6904–6913, 2017. 5, 17

[35] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. BYOL: Bootstrap your own latent: A new approach to self-supervised Learning. *arXiv:2006.07733 [cs, stat]*, Sept. 2020. arXiv: 2006.07733. 2

[36] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross B. Girshick. Masked autoencoders are scalable vision learners. *CoRR*, abs/2111.06377, 2021. 1, 8

[37] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. MAE: Masked Autoencoders Are Scalable Vision Learners. *arXiv:2111.06377 [cs]*, Nov. 2021. arXiv: 2111.06377. 1, 2, 3, 4, 5

[38] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. MoCov1: Momentum Contrast for Unsupervised Visual Representation Learning. pages 9729–9738, 2020. 2

[39] Shell Xu Hu, Da Li, Jan Stühmer, Minyoung Kim, and Timothy M. Hospedales. Pushing the limits of simple pipelines for few-shot learning: External data and fine-tuning make a difference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9068–9077, June 2022. 1

[40] Po-Yao Huang, Hu Xu, Juncheng Li, Alexei Baevski, Michael Auli, Wojciech Galuba, Florian Metze, and Christoph Feichtenhofer. Masked autoencoders that listen, 2022. 8

[41] Zhicheng Huang, Zhaoyang Zeng, Yupan Huang, Bei Liu, Dongmei Fu, and Jianlong Fu. Seeing Out of the Box: End-to-End Pre-Training for Vision-Language Representation Learning. pages 12976–12985, Apr. 2021. 2

[42] Drew A. Hudson and Christopher D. Manning. GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering. pages 6700–6709, 2019. 5, 17

[43] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. *CoRR*, abs/2102.05918, 2021. 1, 4, 8

[44] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. ALIGN: Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. *arXiv:2102.05918 [cs]*, June 2021. arXiv: 2102.05918. 3, 14

[45] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1988–1997, 2017. 5, 7, 17

[46] Kaggle and EyePacs. Kaggle diabetic retinopathy detection. `https://www.kaggle.com/c/diabetic-retinopathy-detection/data`, July 2015. 14

[47] Wonjae Kim, Bokyung Son, and Ildoo Kim. ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision. *arXiv:2102.03334 [cs, stat]*, June 2021. arXiv: 2102.03334. 2

[48] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. Technical Report arXiv:1412.6980, arXiv, Jan. 2017. arXiv:1412.6980 [cs] type: article. 15

[49] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 8

[50] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. *arXiv:1910.13461 [cs, stat]*, Oct. 2019. arXiv: 1910.13461. 2, 4

[51] Chenliang Li, Ming Yan, Haiyang Xu, Fuli Luo, Wei Wang, Bin Bi, and Songfang Huang. SemVLP: Vision-Language Pre-training by Aligning Semantics at Multiple Levels. *arXiv:2103.07829 [cs]*, Mar. 2021. arXiv: 2103.07829. 2

[52] Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Deepak Gotmare, Shafiq Joty, Caiming Xiong, and Steven Hoi. Align before Fuse: Vision and Language Representation Learning with Momentum Distillation. *arXiv:2107.07651 [cs]*, Oct. 2021. arXiv: 2107.07651. 2, 4

[53] Wei Li, Can Gao, Guocheng Niu, Xinyan Xiao, Hao Liu, Jiachen Liu, Hua Wu, and Haifeng Wang. UNIMO-2: End-to-End Unified Vision-Language Grounded Learning. *arXiv:2203.09067 [cs]*, Mar. 2022. arXiv: 2203.09067. 2

[54] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, Lecture Notes in Computer Science, pages 121–137, Cham, July 2020. Springer International Publishing. 5

[55] Yanghao Li, Haoqi Fan, Ronghang Hu, Christoph Feichtenhofer, and Kaiming He. Scaling language-image pre-training via masking, 2022. 2

[56] Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. *CoRR*, abs/2110.05208, 2021. 1

[57] Feng Liang, Yangguang Li, and Diana Marculescu. SupMAE: Supervised Masked Autoencoders Are Efficient Vision Learners. Technical Report arXiv:2205.14540, arXiv, May 2022. arXiv:2205.14540 [cs] type: article. 3

[58] 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*, pages 740–755. Springer, 2014. 16, 19

[59] Ilya Loshchilov and Frank Hutter. SGDR: Stochastic Gradient Descent with Warm Restarts. *arXiv:1608.03983 [cs, math]*, May 2017. arXiv: 1608.03983. 5, 15

[60] Ilya Loshchilov and Frank Hutter. AdamW: Decoupled Weight Decay Regularization. *arXiv:1711.05101 [cs, math]*, Jan. 2019. arXiv: 1711.05101. 5, 15

[61] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. *arXiv:1908.02265 [cs]*, Aug. 2019. arXiv: 1908.02265. 2

[62] Shlok Mishra, Joshua Robinson, Huiwen Chang, David Jacobs, Aaron Sarna, Aaron Maschinot, and Dilip Krishnan. A simple, efficient and scalable contrastive masked autoencoder for learning visual representations, 2022. 8

[63] Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. SLIP: Self-supervision meets Language-Image Pre-training. *arXiv:2112.12750 [cs]*, Dec. 2021. arXiv: 2112.12750. 1

[64] Norman Mu, Alexander Kirillov, David A. Wagner, and Saining Xie. SLIP: self-supervision meets language-image pre-training. *CoRR*, abs/2112.12750, 2021. 1, 2, 5

[65] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation Learning with Contrastive Predictive Coding. *arXiv:1807.03748 [cs, stat]*, Jan. 2019. arXiv: 1807.03748. 2, 3

[66] Hieu Pham, Zihang Dai, Golnaz Ghiasi, Hanxiao Liu, Adams Wei Yu, Minh-Thang Luong, Mingxing Tan, and Quoc V. Le. Combined scaling for zero-shot transfer learning. *CoRR*, abs/2111.10050, 2021. 1, 8

[67] Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models. pages 2641–2649, 2015. 16

[68] 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. CLIP: Learning Transferable Visual Models From Natural Language Supervision. *arXiv:2103.00020 [cs]*, Feb. 2021. arXiv: 2103.00020. 1, 2, 3, 14, 15

[69] 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. *CoRR*, abs/2103.00020, 2021. 1, 5, 8

[70] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. GPT-2: Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 4, 15

[71] Kanchana Ranasinghe, Brandon McKinzie, Sachin Ravi, Yinfei Yang, Alexander T Toshev, and Jonathon Shlens. Perceptual grouping in vision-language models, 2023. 3, 6, 8

[72] Ali Razavi, Aaron van den Oord, and Oriol Vinyals. Generating Diverse High-Fidelity Images with VQ-VAE-2. Technical Report arXiv:1906.00446, arXiv, June 2019. arXiv:1906.00446 [cs, stat] type: article. 2

[73] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: An open large-scale dataset for training next generation image-text models, 2022. 1

[74] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo

Coombes, Jenia Jitsev, and Aran Komatsuzaki. LAION-400M: Open Dataset of CLIP-Filtered 400 Million Image-Text Pairs. *arXiv:2111.02114 [cs]*, Nov. 2021. arXiv: 2111.02114. 6, 14

[75] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. *International Journal of Computer Vision*, 128(2):336–359, Feb. 2020. arXiv:1610.02391 [cs]. 8

[76] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of ACL*, 2018. 14

[77] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. GCC: Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia, 2018. Association for Computational Linguistics. 5, 6

[78] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. FLAVA: A foundational language and vision alignment model. *CoRR*, abs/2112.04482, 2021. 1, 4

[79] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. FLAVA: A Foundational Language And Vision Alignment Model. *arXiv:2112.04482 [cs]*, Feb. 2022. arXiv: 2112.04482. 3

[80] Mannat Singh, Laura Gustafson, Aaron Adcock, Vinicius de Freitas Reis, Bugra Gedik, Raj Prateek Kosaraju, Dhruv Mahajan, Ross B. Girshick, Piotr Dollár, and Laurens van der Maaten. Revisiting weakly supervised pre-training of visual perception models. *CoRR*, abs/2201.08371, 2022. 1

[81] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. VL-BERT: Pre-training of Generic Visual-Linguistic Representations. *arXiv:1908.08530 [cs]*, Feb. 2020. arXiv: 1908.08530. 2

[82] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 1

[83] Yihong Sun, Adam Kortylewski, and Alan Yuille. Amodal Segmentation through Out-of-Task and Out-of-Distribution Generalization with a Bayesian Model. Technical Report arXiv:2010.13175, arXiv, July 2022. arXiv:2010.13175 [cs] type: article. 16

[84] Hao Tan and Mohit Bansal. LXMERT: Learning Cross-Modality Encoder Representations from Transformers. *arXiv:1908.07490 [cs]*, Dec. 2019. arXiv: 1908.07490. 2

[85] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11), 2008. 17

[86] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, \Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 4

[87] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning - ICML '08*, pages 1096–1103, Helsinki, Finland, 2008. ACM Press. 2

[88] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, and Furu Wei. Image as a foreign language: Beit pretraining for all vision and vision-language tasks, 2022. 1

[89] Xiaolong Wang and Abhinav Gupta. Unsupervised Learning of Visual Representations using Videos. Technical Report arXiv:1505.00687, arXiv, Oct. 2015. arXiv:1505.00687 [cs] type: article. 2

[90] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. SimVLM: Simple Visual Language Model Pretraining with Weak Supervision. *arXiv:2108.10904 [cs]*, Aug. 2021. arXiv: 2108.10904. 2

[91] Longhui Wei, Lingxi Xie, Wengang Zhou, Houqiang Li, and Qi Tian. MVP: Multimodality-guided Visual Pre-training. *arXiv:2203.05175 [cs]*, Mar. 2022. arXiv: 2203.05175. 3

[92] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3485–3492. IEEE, 2010. 14

[93] Tete Xiao, Mannat Singh, Eric Mintun, Trevor Darrell, Piotr Dollár, and Ross Girshick. Early Convolutions Help Transformers See Better. Technical Report arXiv:2106.14881, arXiv, Oct. 2021. arXiv:2106.14881 [cs] type: article. 5

[94] Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tieyan Liu. Pre-layernorm: On Layer Normalization in the Transformer Architecture. In *Proceedings of the 37th International Conference on Machine Learning*, pages 10524–10533. PMLR, Nov. 2020. ISSN: 2640-3498. 4, 15

[95] Jianwei Yang, Chunyuan Li, Xiyang Dai, Lu Yuan, and Jianfeng Gao. Focal modulation networks, 2022. 1

[96] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. FILIP: Fine-grained Interactive Language-Image Pre-Training. *arXiv:2111.07783 [cs]*, Nov. 2021. arXiv: 2111.07783. 2

[97] Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. ERNIE-ViL: Knowledge Enhanced Vision-Language Representations Through Scene Graph. June 2020. 2

[98] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models, 2022. 1

[99] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. CoCa: Contrastive Captioners are Image-Text Foundation Models. *arXiv:2205.01917 [cs]*, May 2022. arXiv: 2205.01917. 2

[100] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, Ce Liu, Mengchen Liu, Zicheng Liu, Yumao Lu, Yu Shi, Lijuan Wang, Jianfeng Wang, Bin Xiao, Zhen Xiao, Jianwei Yang, Michael Zeng, Luowei Zhou, and Pengchuan Zhang. Florence: A New Foundation Model for Computer Vision. *arXiv:2111.11432 [cs]*, Nov. 2021. arXiv: 2111.11432. 2

[101] Sukmin Yun, Hankook Lee, Jaehyung Kim, and Jinwoo Shin. Patch-level Representation Learning for Self-supervised Vision Transformers. Technical Report arXiv:2206.07990, arXiv, June 2022. arXiv:2206.07990 [cs] type: article. 2

[102] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow Twins: Self-Supervised Learning via Redundancy Reduction. *arXiv:2103.03230 [cs, q-bio]*, June 2021. arXiv: 2103.03230. 2

[103] Yan Zeng, Xinsong Zhang, and Hang Li. X:VLM: Multi-Grained Vision Language Pre-Training: Aligning Texts with Visual Concepts. *arXiv:2111.08276 [cs]*, Feb. 2022. arXiv: 2111.08276. 3

[104] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12104–12113, June 2022. 3

[105] Xiaohua Zhai, Joan Puigcerver, Alexander Kolesnikov, Pierre Ruyssen, Carlos Riquelme, Mario Lucic, Josip Djolonga, André Susano Pinto, Maxim Neumann, Alexey Dosovitskiy, Lucas Beyer, Olivier Bachem, Michael Tschannen, Marcin Michalski, Olivier Bousquet, Sylvain Gelly, and Neil Houlsby. The visual task adaptation benchmark. *CoRR*, abs/1910.04867, 2019. 1, 14

[106] Xiaohua Zhai, Joan Puigcerver, Alexander Kolesnikov, Pierre Ruyssen, Carlos Riquelme, Mario Lucic, Josip Djolonga, Andre Susano Pinto, Maxim Neumann, Alexey Dosovitskiy, Lucas Beyer, Olivier Bachem, Michael Tschannen, Marcin Michalski, Olivier Bousquet, Sylvain Gelly, and Neil Houlsby. VTAB: A Large-scale Study of Representation Learning with the Visual Task Adaptation Benchmark. *arXiv:1910.04867 [cs, stat]*, Feb. 2020. arXiv: 1910.04867. 5, 7, 14

[107] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. ADE20K: Semantic Understanding of Scenes through the ADE20K Dataset. *arXiv:1608.05442 [cs]*, Oct. 2018. arXiv: 1608.05442. 19