SLIP: Self-supervision meets Language-Image Pre-training

Norman Mu[1], Alexander Kirillov[2], David Wagner[1], and Saining Xie[2]

[1] UC Berkeley
[2] Meta AI

Abstract Recent work has shown that self-supervised pre-training leads to improvements over supervised learning on challenging visual recognition tasks. CLIP, an exciting new approach to learning with language supervision, demonstrates promising performance on a wide variety of benchmarks. In this work, we explore whether self-supervised learning can aid in the use of language supervision for visual representation learning with Vision Transformers. We introduce SLIP, a multi-task learning framework for combining self-supervised learning and CLIP pre-training. After pre-training, we thoroughly evaluate representation quality and compare performance to both CLIP and self-supervised learning under three distinct settings: zero-shot transfer, linear classification, and end-to-end finetuning. Across ImageNet and a battery of additional datasets, we find that SLIP improves accuracy by a large margin. We validate our results further with experiments on different model sizes, training schedules, and pre-training datasets. Our findings show that SLIP enjoys the best of both worlds: better performance than self-supervision (+8.1% linear accuracy) and language supervision (+5.2% zero-shot accuracy). Our code is available at github.com/facebookresearch/SLIP

1 Introduction

Much of recent progress in deep learning has been driven by the paradigm of pre-training powerful, general-purpose representations that transfer well to a variety of specific applications. Within computer vision, supervised learning on image classification and self-supervised learning on unlabeled images comprise the two primary approaches to representation learning. After AlexNet [23], researchers soon realized that supervised pre-training yields a generic visual backbone which can be repurposed for many different tasks [13]. Today, most state-of-the-art results still depend on supervised pre-training, and scaling to massive amounts of data, such as Google's proprietary JFT dataset, remains one of the most reliable methods for improving downstream performance. Self-supervised learning, a form of unsupervised learning, found tremendous success first in the domain of language [9,29], but has also made significant recent progress in vision. A major motivation for studying self-supervised learning has been a desire to supersede supervised pre-training and its reliance on labor-intensive human annotation. Indeed, self-supervised pre-training has outperformed supervised learning for some

time now on small datasets, but only recently with the development of contrastive methods [5,18] has it begun to improve performance on larger datasets such as ImageNet.

Both supervised and self-supervised pre-training today rely heavily on ImageNet (*i.e.* ImageNet-1K) [30], a highly curated dataset with particular idiosyncrasies and biases [35]. The YFCC100M dataset [33] was released in 2015 and remains the largest publicly-accessible collection of images. To date, the field of representation learning has found much less use for this dataset. On the other hand, the full ImageNet dataset of 14M images (*i.e.* ImageNet-22K) has become very popular for its role in training Vision Transformer models which require a larger amount of data than ImageNet-1K [10,1]. Why are uncurated datasets not more common in the study of representation learning? There are a few possible reasons. Most immediately, uncurated datasets also lack labels and so long as supervised pre-training remains the simpler and more accessible option for most researchers, datasets like YFCC100M are a non-starter. As we confirm again in our work, the standard self-supervised evaluation task of ImageNet classification from frozen features heavily biases results against models not also pre-trained on ImageNet [2]. Finally, while progress on ImageNet has been encouraging, there has not been strong evidence that current self-supervised methods scale well to larger uncurated datasets [34,11].

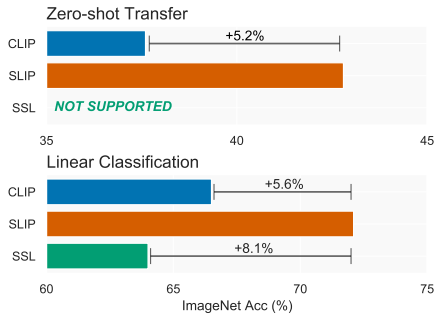

Figure 1: **SLIP pre-training on YFCC15M.** Combining image-only self-supervision and image-text supervision simultaneously improves zero-shot transfer and linear classification on ImageNet.

Recently, CLIP [28] introduced an exciting new approach to representation learning. It re-examines language supervision for learning visual representations, and catapults it into contention with label supervision and self-supervision. CLIP requires only images and free-form text captions, thus revitalizing the use of YFCC100M in representation learning. In addition to no longer requiring label annotations, CLIP accuracy also scales well to large datasets and models. The best results for CLIP are achieved with big models on a curated dataset of 400M image-text pairs, though promising results are also shown on a subset of YFCC100M. CLIP also enables many exciting new applications with its flexible language-guided capabilities.

In this work, we explore whether the momentum of self-supervised learning on images carries into the setting of language supervision. In particular, we

investigate whether language supervision in the form of CLIP also benefits from image self-supervision. We note that it is not immediately clear that jointly training on these two objectives should improve representation quality, since the objectives require the model to encode different and possibly conflicting information about the image.

In order to explore these questions, we introduce SLIP (**S**elf-supervision meets **L**anguage-**I**mage **P**re-training), a multi-task framework combining language supervision and self-supervision. We pre-train various SLIP models on a subset of YFCC100M, and thoroughly evaluate representation quality under three distinct settings: zero-shot transfer, linear classification, and end-to-end finetuning. We evaluate downstream performance on ImageNet, in addition to a battery of 25 other classification benchmarks. Additionally, we further validate our findings with experiments on different model sizes, training schedules, and pre-training datasets. Our findings conclusively show that SLIP improves performance across most evaluations by a significant margin, an encouraging signal for the general utility of self-supervision in the context of language supervision. Additionally, we analyze various components of our method in further detail such as the choices of pre-training dataset and data processing method. We conclude with a discussion of our evaluations.

2 Related Work

Language supervision. Early work explored learning visual representations from image captions, even before the advent of deep learning [27]. DeViSE [12] jointly embeds images and textual class labels within a shared semantic space, allowing the model to recognize classes that were not explicitly trained for. Initial attempts at leveraging the YFCC100M dataset for representation learning included predicting the bag-of-words representation [22] or n-gram occurrence [24] from images. ICMLM [31] and VirTex [8] showed that language supervision on COCO Captions produced useful visual representations. Prior to CLIP, Multimodal Contrastive Training [38] adds contrastive image-image and language-image losses to VirTex which further improve performance. CLIP [28] quickly garnered significant attention for its simplicity, scale, and strong results. Developed concurrently, ALIGN [21], uses a larger but noisier uncurated dataset and shows similar results.

Self-supervised learning. Earlier self-supervised learning methods have shown subpar scaling with dataset size [15]. Contrastive learning methods ushered in rapid progress [26,37,18,5] due to their simplicity and effectiveness. Recent methods for self-supervised learning also propose a variety of alternatives to the contrastive objective such as self-distillation [16,3], or input reconstruction [1,17].

Multi-modal multi-task learning. MURAL [20] extends ALIGN to the multi-lingual setting and introduces a cross-lingual objective to improve multi-lingual image and text retrieval. Concurrently to this work, DeCLIP [25] adds several additional training objectives and more data collected in-house to CLIP in order to improve data efficiency.

3 SLIP Framework

We introduce SLIP, a framework for combining language supervision and image self-supervision to learn visual representations without category labels. During pre-training, separate views of each input image are constructed for the language supervision and image self-supervision branches, then fed through a shared image encoder. Through the course of training, the image encoder learns to represent visual input in a semantically meaningful manner. We then measure the quality of these representations through performance on downstream tasks.

3.1 Contrastive Language-Image Pre-training

Radford et al. [28] demonstrated the ability of contrastive learning (CLIP) on corresponding images and captions to learn powerful representations. CLIP embeds images and text with separate modality-specific models. These vectors are then projected into a shared embedding space and normalized. The InfoNCE loss is computed using these embeddings, with corresponding images and captions as positive pairs and all non-matching images and captions as negative pairs.

Non-contrastive alternatives for language supervision include predicting a bag-of-words representation of the caption [22] or the original caption [31,8] from the image. However, the authors of [29] find that these methods to be less effective than CLIP. The contrastive objective also enables image classification without re-training (zero-shot transfer).

3.2 Image Self-Supervision

View-based self-supervised learning, in which models are trained to represent different views or augmentations of the same image similarly, has yielded strong results across a variety of different formulations. In this work we primarily use an adaptation of SimCLR [5,6], a representative example of these methods, as the self-supervised objective in SLIP. However, other frameworks can be swapped in quite easily, and we explore this in Section 6. We focus on the Vision Transformer [10] architecture for its simplicity and good performance. We follow hyperparameter settings from MoCo v3 [7] for training self-supervised Vision Transformers, which will be described later in Section 4.1.

3.3 Our Method

We outline SLIP with SimCLR for self-supervision (*i.e.* SLIP-SimCLR). The pseudo-code for our algorithm can be found in the appendix. During each forward pass in SLIP, all images are fed through the image encoder. The CLIP and SSL objectives are computed on the relevant embeddings and then summed together into a single scalar loss. The two objectives can be balanced differently by rescaling the SSL objective. We find that a scale of 1.0 for the self-supervised objective, i.e. no re-scaling, works well for SimCLR. Unless otherwise noted, we refer to SLIP-SimCLR simply as SLIP.

SLIP increases the number of images processed which results in approximately 3× more activations. This expands the model's memory footprint and slows down the forward pass during training. See Section 7 for further discussion.

4 Improved Training Procedure

The authors of CLIP focus primarily on training with a large private dataset of 400M image-text pairs, where the large scale of data lessens the need for regularization and data augmentation. While re-implementing CLIP, we found some simple adjustments (mostly to data processing) which significantly improved model performance when pre-trained on YFCC15M. Our improved training procedure, detailed in the appendix, achieves 34.6% zero-shot transfer to ImageNet with a modified[3] ResNet-50, exceeding the original result of 31.3% [28]. Another re-implementation achieves 32.7% accuracy on ImageNet [19]. In our experiments we focus primarily on the Vision Transformer model family for their strong scaling behavior [10]. We train all Vision Transformer models with our improved procedure as well, in order to set strong baselines for comparing our methods.

4.1 Implementation Details

Datasets. We focus primarily on a 15M subset of YFCC100M [33] filtered by Radford et al. [28] consisting of English-only titles and descriptions, which we refer to as YFCC15M. We also evaluate on Conceptual Captions 3M (CC3M) [32] and Conceptual Captions 12M (CC12M) [4].
Data Augmentation. During training, we randomly sample a valid caption for each image (i.e. title or description for YFCC15M). Images for the CLIP branch are randomly resized and cropped to between 50% and 100% of the original image, which we refer to as global cropping. In the self-supervised branch we sample two views with the augmentation from MoCo v3 [5].
Architecture. We use the original ViT-B/16 and ViT-L/16 architectures from the ViT paper [10] for our image encoders, as well as a ViT-S/16 architecture [36] which is comparable to ResNet-50 in FLOPs and parameters. For our text encoders, we use the smallest text Transformer model from CLIP which contains 63M parameters and uses byte-pair encoding with a 49K token vocabulary, and maximum context length of 77.

For the CLIP objective, our model projects the image and caption embeddings into a 512-dim space with separate learned linear projections. In the self-supervised branch, we use the 3-layer MLP projection head with 4096-dim hidden layers to transform the image embeddings into a 256-dim output space.
Training. We train with a batch size of 4096 and the AdamW optimizer in all our experiments. Both the image and text encoders are randomly initialized. Following CLIP, we set the $\beta_2 = 0.98$ to improve training stability, but we keep

[3] The initial 7×7 conv is replaced by three 3×3 convs; global average pooling is replaced by a self-attention pooling layer with 14M parameters.

$\epsilon = 1e-8$. We use a weight decay of 0.5 for CLIP and 0.1 for SLIP. Instead of the custom mixed-precision recipe used in CLIP, we opt for the built-in automatic mixed precision library in PyTorch.

Zero-shot Transfer Evaluation. We evaluate zero-shot transfer to various classification benchmarks including ImageNet. We perform prompt ensembling by averaging the caption embeddings for each class across the prompt templates. This average caption embedding is then used to compute cosine similarity with the image embeddings. CLIP provides prompt templates and class names for these benchmarks, which we use directly for ease of comparison.

Linear Classification Evaluation. We use the same setup as MoCo v3 to evaluate linear classification performance. We use SGD w/ momentum and no weight decay. On ImageNet, we use a learning rate of 0.01 and on the other downstream datasets we tune the learning rate and report the best result. We train for 100 epochs and perform standard cropping and flipping augmentations.

End-to-end Finetuning Evaluation. To finetune our models on ImageNet, we use the training procedure from BEiT [1]. This procedure employs strong regularization and data augmentation, as well as layerwise learning rate decay. We disable relative positional embedding, layer scaling, and average pooling across tokens. For ViT-B and ViT-S we train for 100 epochs, while on ViT-L we train for 50 epochs. For finetuning on smaller downstream datasets, we use the simpler DeiT training procedure [36].

5 Empirical Evaluations

5.1 ImageNet Classification

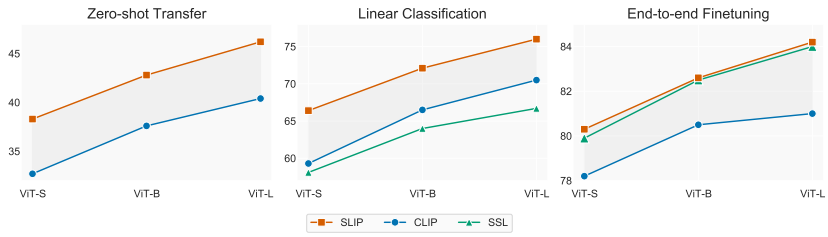

Figure 2: **ImageNet results.** We evaluate the representation quality by testing the performance on ImageNet under different settings: zero-shot transfer using text prompts, linear classification, and end-to-end finetuning. SLIP improves upon the zero-shot transfer and linear classification performance of CLIP by significant margin across all vision Transformer model sizes.

We evaluate performance on ImageNet under three distinct settings: zero-shot transfer, linear classification, and end-to-end finetuning. The zero-shot

transfer task evaluates model performance on classification benchmarks directly after pre-training without updating any of the model weights. A model trained with contrastive language supervision can be used as an image classifier by selecting the class whose caption embedding aligns most closely with the input image. Linear classification, also called linear probing, is a standard evaluation method used to evaluate unsupervised or self-supervised representations. A randomly initialized final classification layer is trained while all other model weights are frozen. Finally, another way of evaluating representation quality is whether a pre-trained model can improve upon the performance of supervised learning when finetuning the model end-to-end.

Dataset	Method	Linear	Finetuning
ImageNet	SimCLR	74.5	82.8
	MoCo v3	76.6	83.1
YFCC15M	SimCLR	64.0 (-10.5)	82.5 (-0.3)
	MoCo v3	66.1 (-10.5)	82.8 (-0.3)

Table 1: We train ViT-B/16 with two self-supervised frameworks. Both linear classification and end-to-end finetuning accuracy on ImageNet suffers when pre-training on YFCC15M instead of ImageNet. Accuracy drop show in (red).

One common evaluation setup in the self-supervised learning literature is to train both the model and the linear classifier on ImageNet (i.e. ImageNet-1K), which even without labels is a highly curated and class-balanced dataset. In Table 1 we train ViT-B/16 with SimCLR and MoCo v3 on both YFCC15M and ImageNet. The resulting models are evaluated on ImageNet using linear classification and end-to-end finetuning. Both SimCLR and MoCo v3 experience a more than 10% drop in linear classification accuracy when pretrained on YFCC15M instead of ImageNet, a dramatic degradation in performance. For this reason, the baseline linear results in our experiments are lower than what is typically reported in the self-supervised literature. Similarly, we observe a less severe but consistent degradation for end-to-end finetuning results as well. We argue that training on uncurated data is a more realistic and informative setting, especially given the original motivations of learning vision from less supervision.

In Table 2, we provide evaluation results for CLIP, SimCLR, and SLIP across three sizes of Vision Transformer and on all three ImageNet settings. All models are trained for 25 epochs on YFCC15M. We find that language supervision and image self-supervision interact constructively in SLIP, improving upon the performance of both methods alone.

Zero-shot transfer. Self-supervised models do not support zero-shot transfer evaluation since there is no way to directly map the learned representations onto categorical labels. SLIP consistently outperforms CLIP by around +5% on zero-shot transfer across all three model sizes, a very large margin relative to the original number. The gap between SLIP and CLIP does close slightly

Model	Method	0-shot	Linear	Finetuned
ViT-S/16	CLIP	32.7	59.3	78.2
	SimCLR	-	58.1	79.9
	SLIP	**38.3** (+5.6)	**66.4** (+7.1)	**80.3** (+0.4)
ViT-B/16	CLIP	37.6	66.5	80.5
	SimCLR	-	64.0	82.5
	SLIP	**42.8** (+5.2)	**72.1** (+5.6)	**82.6** (+0.1)
ViT-L/16	CLIP	40.4	70.5	81.0
	SimCLR	-	66.7	84.0
	SLIP	**46.2** (+4.8)	**76.0** (+5.5)	**84.2** (+0.2)

Table 2: Full ImageNet results. SLIP significantly improves performance on ImageNet in the zero-shot transfer, linear classification, and end-to-end finetuning settings. Improvements over stronger baseline (underlined) shown in (green).

between ViT-Small (22M params) and ViT-Large (300M params) from +5.6% to +4.8%. This trend suggests that SLIP would continue to yield benefits over CLIP even for the largest Vision Transformer architectures currently in use. With ViT-Large SLIP achieves 46.2% top-1 accuracy, which is still significantly below the performance achieved with larger curated data [28]. In absolute terms however, this is a surprisingly strong result considering that YFCC15M contains very little data of the specific form seen during zero-shot transfer evaluation (i.e. object-centric, or iconic, images labeled with captions of the form "a photo of a class name.").

Linear Classification. In this setting we also observe the benefits of combining language supervision and image self-supervision. CLIP outperforms SimCLR, but by a much smaller margin than SLIP outperforms SimCLR. We see that SLIP significantly outperforms SimCLR in linear classification accuracy across all three model sizes. The gap between SLIP and SimCLR is largest with ViT-L at almost +10%, suggesting that SLIP continues to scale with larger models while SimCLR slightly saturates in performance.

End-to-end Finetuning. We see in Table 1 that finetuning performance is somewhat less affected by pre-training on YFCC15M than linear performance is affected, possibly because the model is allowed to adapt to the target distribution. Both SimCLR and MoCo v3 experience -0.3% drops in finetuning accuracy when pre-trained on YFCC15M instead of ImageNet, which is still quite significant for this setting. We re-iterate that the results in Table 2 are not directly comparable with methods which are pre-trained on ImageNet-1K.

When finetuning on ImageNet, CLIP is particularly weak: ViT-S and ViT-B performance is below even that of training from a random weight initialization [36]. The performance of CLIP does not scale well with model size either, as CLIP ViT-L performance is only +0.5% above CLIP ViT-B. On the other hand, self-supervised learning does quite well in this setting, especially with the larger models. SimCLR ViT-L enjoys a +3.0% gain in accuracy over CLIP ViT-L, and

SLIP ViT-L does slightly better than SimCLR ViT-L, though by a very marginal amount. These results suggest that the low finetuning performance of CLIP is mostly solved with self-supervision.

5.2 Model and Compute Scaling

We also investigate the scaling behavior of SLIP with more compute (longer training) and larger vision models. We note that 100 epochs of training on YFCC15M corresponds to around 1200 epochs of training on ImageNet-1K. In Table 3 we experimented with holding model size fixed (ViT-B/16) and training for longer as well as training different model sizes for an extended training schedule (100 epochs). Our results indicate that SLIP scales well with both longer training and larger models. We show full results simultaneously varying model and compute scaling with SLIP in the appendix.

Model	#params.	0-shot	Linear	Finetuned
ViT-S	22M	39.5	68.3	80.7
ViT-B	86M	45.0	73.6	83.4
ViT-L	307M	47.9	75.1	84.8

(a) Comparing ViT model variants of different capacities (ViT-S/B/L). All models are pre-trained for 100 epochs.

Epochs	0-shot	Linear	Finetuned
25	42.8	72.1	82.6
50	44.1	73.0	82.9
100	45.0	73.6	83.4

(b) ViT-B with longer pre-training schedules (25/50/100 epochs).

Table 3: SLIP pre-training performance (in terms of zero-shot transfer, linear classification, and end-to-end finetuning) can scale well with both model size and number of training epochs.

5.3 Additional Benchmarks

		Food-101	CIFAR-10	CIFAR-100	CUB	SUN397	Cars	Aircraft	DTD	Pets	Caltech-101	Flowers	MNIST	FER-2013	STL-10	EuroSAT	RESISC45	GTSRB	KITTI	Country211	PCAM	UCF101	Kinetics700	CLEVR	HatefulMemes	SST2	ImageNet	Average
ViT-S	CLIP	43.4	61.0	29.9	31.1	43.9	3.1	4.7	17.9	25.0	53.3	47.8	9.8	29.1	86.8	22.3	16.1	9.5	34.1	8.7	**64.8**	26.0	18.8	14.7	**56.1**	49.5	32.7	32.3
	SLIP (25 ep)	51.6	**73.0**	35.4	36.3	49.2	4.2	**6.1**	25.7	30.9	62.8	54.3	**9.9**	31.3	91.6	**22.4**	21.9	11.0	**39.9**	**9.6**	50.8	32.8	22.9	14.8	49.6	**50.1**	38.3	35.6
	SLIP (100 ep)	**53.0**	68.4	**39.3**	**36.5**	**49.8**	**4.6**	5.1	**26.6**	**33.6**	**68.3**	**55.8**	2.7	**37.8**	**91.9**	18.2	**22.2**	**13.8**	38.4	8.5	62.8	**33.3**	**23.5**	**19.2**	51.4	49.4	**39.5**	**36.7**
ViT-B	CLIP	50.6	66.0	34.5	38.8	51.1	4.0	5.4	21.2	28.5	60.9	53.3	8.4	17.3	90.5	**30.2**	21.5	6.1	35.1	10.5	53.5	28.5	22.1	10.8	52.4	50.7	37.6	34.2
	SLIP (25 ep)	59.5	78.6	45.2	38.7	**53.4**	5.4	5.7	26.1	31.1	71.0	56.6	9.8	19.6	94.4	20.3	**28.9**	**14.5**	34.0	**11.6**	**55.4**	**37.7**	**26.9**	**17.5**	52.8	**51.1**	42.8	38.0
	SLIP (100 ep)	**63.3**	**79.2**	**50.4**	**44.7**	52.0	**8.1**	**8.4**	**26.2**	**34.7**	**74.0**	**61.3**	**17.1**	**40.8**	**95.4**	20.8	27.8	11.7	**35.2**	11.5	52.1	37.1	25.8	13.0	**55.1**	49.9	**45.0**	**40.0**
ViT-L	CLIP	59.5	72.9	41.5	**40.3**	53.6	6.9	6.4	20.6	27.9	65.4	55.0	10.3	34.5	94.2	22.7	28.8	5.8	**41.4**	12.6	**54.9**	34.3	24.0	12.9	54.3	**50.1**	40.4	37.4
	SLIP (25 ep)	64.4	**87.8**	**56.4**	39.8	**58.9**	8.6	7.8	26.8	32.0	76.6	59.4	13.2	36.0	**96.6**	27.7	36.5	7.2	28.8	**15.6**	54.4	42.6	30.0	14.1	53.4	**50.1**	46.2	41.2
	SLIP (100 ep)	**69.2**	87.5	54.2	39.8	56.0	**9.0**	**9.5**	**29.9**	**41.6**	**80.9**	**60.2**	**14.9**	**39.6**	96.2	**34.5**	**46.0**	**8.6**	30.7	14.2	50.6	**44.1**	**30.5**	**17.4**	**55.0**	49.8	**47.9**	**43.0**

Table 4: Zero-shot transfer evaluation with ViT S, B, and L on a variety of classification benchmarks. Best results in **bold**. SLIP outperforms CLIP on most of the tasks, frequently with a significant margin. With longer pre-training epochs, the performance can be further improved.

While evaluating classification performance on ImageNet gives a broad overview of representation quality, it is also informative to measure performance on a variety of narrowly targeted downstream datasets. In Table 4 we evaluate zero-shot transfer on a battery of downstream image classification tasks compiled by [28]. We also provide linear classification results on these benchmarks in the appendix. These datasets span many different domains including everyday scenes such as traffic signs, specialized domains such as medical and satellite imagery, video frames, rendered text with and without visual context, and more. We remove Pascal VOC and replace NABirds with CUB-200-2011. To preprocess the datasets into a unified pipeline we use the extra scripts included in VISSL [14]. We catalog chance performance along with short descriptions of the datasets in the appendix.

In the zero-shot setting, both CLIP and SLIP models perform well on datasets whose categories are well represented in YFCC15M, such as Food-101, Oxford Pets, Caltech-101, and STL-10. On these datasets we see that larger models and training for longer with SLIP more generally improve zero-shot transfer accuracy. Datasets with less overlap with the content in YFCC15M, such as Rendered SST2, KITTI depth, and PatchCamelyon (PCAM) is only around chance performance.

Zero-shot performance on the low-resolution datasets (MNIST, CIFAR-10, CIFAR-100) is also very poor. On many datasets performance is several times chance performance yet still much lower than what is achievable with a small supervised model. This suggests that language supervision alone is an inefficient way of training models for specific tasks of interest. Which method does best under the zero-shot setting is also somewhat inconsistent and variable across datasets, unlike the linear setting, and we note this as a caveat in evaluating representation quality with zero-shot evaluations.

5.4 Additional Pre-training Datasets

Dataset	Method	0-shot	Linear	Finetuned
CC3M	CLIP	17.1	53.3	79.5
	SimCLR	-	55.4	80.9
	SLIP	23.0	65.4	81.4
CC12M	CLIP	36.5	69.0	82.1
	SimCLR	-	62.2	82.6
	SLIP	40.7	73.7	83.1

Table 5: ImageNet results with ViT-B/16 pre-trained on CC3M [32] and CC12M [4], two smaller datasets.

In addition to YFCC15M, we experiment with two additional image-text datasets: CC3M and CC12M. In Table 5, we train ViT-B with CLIP, SimCLR,

and SLIP. SLIP maintains its margin of improvement over CLIP and SimCLR in all ImageNet evaluation settings. Notably, pre-training SLIP on CC12M instead of YCC15M yields lower zero-shot accuracy but results in higher linear and finetuning performance. CLIP sees a boost to finetuning performance of +1.6%.

Our improved training recipe (see Section 4.1) alleviates overfitting by CLIP when trained on YFCC15M and CC12M, but on the smaller CC3M dataset CLIP overfits quite dramatically. This may also be due to the hypernymization and other aggressive text cleaning used in CC3M to make the captions more amenable to image captioning but reduces dataset difficulty. CLIP reaches its highest zero-shot ImageNet accuracy after just 15 out of 40 epochs of training on CC3M, after which we observe a steady decline in ImageNet zero-shot transfer accuracy. In contrast, on CC3M SLIP reaches its highest zero-shot ImageNet performance after 35 epochs.

5.5 Alternative Self-Supervised Frameworks

Method	0-shot	Linear	Finetuned
SLIP-SimCLR [5]	42.8	72.1	82.6
SLIP-MoCo v3 [7]	41.8	71.4	82.4
SLIP-BYOL [16]	41.3	71.1	82.2
SLIP-BEiT [1]	39.1	66.5	82.2
None (CLIP)	37.6	66.5	80.5

Table 6: We evaluate ViT-B/16 with several SLIP variants using different self-supervised frameworks. SLIP works the best with SimCLR among several other self-supervised frameworks, but all variants outperform CLIP.

As noted in Section 3.2, SLIP enables the use of many different self-supervision methods. We ran several experiments on ViT-B/16 with different alternatives to SimCLR, in particular MoCo v3 [7], BYOL [16], and BEiT [1]. Similar to how we tuned the hyperparameters for SLIP-SimCLR, we largely keep the original self-supervised hyperparameters and add in the CLIP objective and text encoder. MoCo v3 and BEiT hyperparameters are already tuned for ViT, but with BYOL we tuned the learning rate and weight decay while copying the data augmentation and projector/predictor heads from MoCo v3. We also lightly tune different scaling parameters for the self-supervised loss. All models are trained for 25 epochs on YFCC15M.

Results in Table 6 show that all three alternatives underperform SLIP-SimCLR, despite being individually stronger self-supervised methods. Nonetheless, all SLIP variants still improve performance over CLIP.

6 Further Analysis

What do language supervised models learn from YFCC15M? We probe the sources of the image classification abilities of CLIP and SLIP by visualizing nearest neighbor retrievals from the YFCC15M training data using each model's image encoder, shown in the appendix. Our visualizations reveal a surprising amount of specific and accurate category information in the captions (object names, plant and animal species, geographic location, etc).

We also estimate an upper bound of zero-shot ImageNet classification performance using YFCC15M with a simple image retrieval baseline. With a strong ImageNet classifier (BEiT-Large @ 384px[4]), we retrieve the 50 nearest neighbors of each validation image from the YFCC15M training images. We then map each caption of the retrieved images onto ImageNet classes by selecting the closest class text embedding as measured by the publicly released CLIP ViT-L/16 text encoder trained on 400M image-text pairs. We take the modal class as the classification prediction. Thus, each validation image can only be correctly classified if there exists similar training images in YFCC15M which are captioned in a way that describes the correct ImageNet category. This baseline achieves surprisingly high 74.4% top-1 accuracy, indicating a substantial amount of accurate, category-specific information in the captions.

What does SLIP gain from self-supervision? We evaluate the image retrieval baseline from above using the image encoders from our SLIP and CLIP models, shown in Table 7. We also measure the average cosine similarity between ImageNet image embeddings (averaged across 50 validation images per class) and the corresponding class embedding (averaged across 7 prompts) for these two models, and find much higher similarity between image and text for SLIP than CLIP. We interpret these two results to support the conclusion that the self-supervision objective pushes SLIP to learn better visual features, which are then more easily indexed by the text encoder.

Method	CLIP	SLIP
image retrieval acc.	26.3%	**29.1%**
cosine similarity	0.343	**0.412**

Table 7: Comparison of SLIP vs. CLIP feature quality with a image retrieval baseline and average cosine similarity between images and categories on ImageNet. Both methods use a ViT-B/16 model trained on YFCC15M.

Why not pre-train with SSL and finetune with CLIP? An alternative to SLIP would be to simply initialize the image encoder of CLIP with SSL-trained weights. We try training CLIP ViT-B/16 under this setting but find worse performance than training jointly with CLIP and SSL. In 8, we see this approach underperforms SLIP in all three ImageNet evaluation settings.

[4] This model achieves 88.4% top-1 accuracy on ImageNet.

Method	0-shot	Linear	Finetuned
SimCLR → CLIP	41.1	68.2	82.3
SLIP-SimCLR	**42.8**	**72.1**	**82.6**

Table 8: Finetuning vs. multi-task training. One alternative to SLIP consists of initializing the image encoder of CLIP with weights trained through self-supervised learning. With ViT-B/16 trained for 25 epochs, finetuning with CLIP performs noticeably worse across all three ImageNet evaluating settings.

Is SLIP just CLIP with data augmentation? We also examine the effects of adding further data augmentation to CLIP and whether this explains the performance improvements seen in SLIP. The SimCLR augmentation can be separated into two components: color (jitter or grayscale) + blur, and resize crop + flip. We train CLIP with these two components individually and also with the full SimCLR augmentation. When training with color + blur, we use the original CLIP cropping strategy from [28] in which we resize the shorter side to 224px then perform a random square crop. Our results are shown in Table 9. While augmentation and resize crop + flip hurt performance, color + blur do improve zero-shot transfer performance by +0.8% which is still far below the gain by SLIP.

Augmentation	0-shot	Linear	Finetuned
global crop (CLIP)	37.6	66.5	80.5
color + blur	38.4	68.5	81.5
resize crop + flip	36.0	66.1	80.5
color + blur + resize crop + flip	36.3	65.2	80.6
SLIP	**42.8**	**72.1**	**82.6**

Table 9: We train CLIP with different data augmentations and compare ImageNet performance to SLIP. Color + blur slightly improve performance over our improved training recipe using global image crops, but by a much smaller margin than SLIP does.

Can we fully decouple self-supervision from language supervision? We experiment with a version of SLIP we call SLIP-decoupled in which the self-supervised objective is computed on a disjoint set of 15M images from the YFCC15M images used in the text supervision object. During training, the images are sampled independently from both sets, effectively decoupling the language-image supervision and self-supervision signals. In Table 10, we find that SLIP-decoupled does just as well as SLIP.

Method	0-shot	Linear	Finetuned
SLIP	42.8	72.1	82.6
Decoupled SLIP	42.7	72.0	82.8

Table 10: Decoupling self-supervision and text-supervision has no effect on performance. We sampled an additional 15M images disjoint from the YFCC15M images to use only in the self-supervised objective and observe that this performs nearly identically.

7 Discussion

Our results on ImageNet and other classification benchmarks show that language supervision and self-supervision are indeed highly synergistic. As shown in Table 2, SLIP improves zero-shot performance across model sizes by large margins of +4.8% to +5.6%. Similar gains can be seen in the linear classification setting, with consistent but marginal improvements in the end-to-end finetuning setting.

These trends remain consistent on longer training schedules with the exception of linear probe performance on SLIP ViT-L which actually decreases with more training. With SLIP ViT-L pre-trained on YFCC15M for 100 epochs, we achieve our strongest result of 47.9% zero-shot accuracy on ImageNet. SLIP also shows significant improvements on CC3M and CC12M. Finally, we also confirm our findings with zero-shot and linear evaluations on additional benchmarks.

Evaluating representation quality. Prior work on representation learning has argued against end-to-end finetuning for its sensitivity to optimization hyperparameters [15], and against linear classification for being too contrived [39]. We instead view zero-shot transfer, along with linear classification and end-to-end finetuning, as one cohesive paradigm for evaluating representation quality. Zero-shot transfer represents the strictest setting, where the exemplar vector for each class must be specified through natural language. Linear classification is a relaxation of zero-shot transfer, in which the class exemplars are optimized on training data. Finally, end-to-end finetuning represents a further relaxation where all model parameters are allowed to adapt to. Performance should be assessed across multiple settings, rather than a single setting.

Zero-shot ImageNet monitor. SLIP may also serve as a useful framework within which to evaluate new methods for self-supervised learning. Training loss on the pre-text task is a poor predictor of downstream performance, so a simple external metric like kNN accuracy is important for quickly estimating performance and diagnosing training issues such as overfitting or instability. However, kNN classification requires encoding and storing every single training image and naive inference requires very expensive matrix multiplications. The memory bank kNN monitor [7] alleviates this cost but is not feasible when pre-training on unlabeled datasets such as YFCC100M. Instead, zero-shot evaluations on ImageNet are virtually as fast as evaluating validation accuracy in the supervised setting.

Acknowledgements. This work was supported by BAIR, the Berkeley Deep Drive (BDD) project, and gifts from Meta and Open Philanthropy.

References

1. Bao, H., Dong, L., Wei, F.: Beit: Bert pre-training of image transformers. ArXiv **abs/2106.08254** (2021)
2. Caron, M., Bojanowski, P., Joulin, A., Douze, M.: Deep clustering for unsupervised learning of visual features. In: ECCV (2018)
3. Caron, M., Touvron, H., Misra, I., J'egou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. ArXiv **abs/2104.14294** (2021)
4. Changpinyo, S., Sharma, P.K., Ding, N., Soricut, R.: Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 3557–3567 (2021)
5. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.E.: A simple framework for contrastive learning of visual representations. ArXiv **abs/2002.05709** (2020)
6. Chen, T., Kornblith, S., Swersky, K., Norouzi, M., Hinton, G.E.: Big self-supervised models are strong semi-supervised learners. ArXiv **abs/2006.10029** (2020)
7. Chen, X., Xie, S., He, K.: An empirical study of training self-supervised vision transformers. ArXiv **abs/2104.02057** (2021)
8. Desai, K., Johnson, J.: Virtex: Learning visual representations from textual annotations. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 11157–11168 (2021)
9. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: NAACL (2019)
10. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. ArXiv **abs/2010.11929** (2021)
11. El-Nouby, A., Izacard, G., Touvron, H., Laptev, I., Jégou, H., Grave, E.: Are large-scale datasets necessary for self-supervised pre-training? ArXiv **abs/2112.10740** (2021)
12. Frome, A., Corrado, G.S., Shlens, J., Bengio, S., Dean, J., Ranzato, M., Mikolov, T.: Devise: A deep visual-semantic embedding model. In: NIPS (2013)
13. Girshick, R.B., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. 2014 IEEE Conference on Computer Vision and Pattern Recognition pp. 580–587 (2014)
14. Goyal, P., Duval, Q., Reizenstein, J., Leavitt, M., Xu, M., Lefaudeux, B., Singh, M., Reis, V., Caron, M., Bojanowski, P., Joulin, A., Misra, I.: Vissl. https://github.com/facebookresearch/vissl (2021)
15. Goyal, P., Mahajan, D.K., Gupta, A., Misra, I.: Scaling and benchmarking self-supervised visual representation learning. 2019 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 6390–6399 (2019)
16. Grill, J.B., Strub, F., Altch'e, F., Tallec, C., Richemond, P.H., Buchatskaya, E., Doersch, C., Pires, B.Á., Guo, Z.D., Azar, M.G., Piot, B., Kavukcuoglu, K., Munos, R., Valko, M.: Bootstrap your own latent: A new approach to self-supervised learning. ArXiv **abs/2006.07733** (2020)
17. He, K., Chen, X., Xie, S., Li, Y., Doll'ar, P., Girshick, R.B.: Masked autoencoders are scalable vision learners (2021)
18. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.B.: Momentum contrast for unsupervised visual representation learning. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 9726–9735 (2020)

19. Ilharco, G., Wortsman, M., Carlini, N., Taori, R., Dave, A., Shankar, V., Namkoong, H., Miller, J., Hajishirzi, H., Farhadi, A., Schmidt, L.: Openclip (Jul 2021). https://doi.org/10.5281/zenodo.5143773, https://doi.org/10.5281/zenodo.5143773, if you use this software, please cite it as below.
20. Jain, A., Guo, M., Srinivasan, K., Chen, T., Kudugunta, S., Jia, C., Yang, Y., Baldridge, J.: Mural: Multimodal, multitask retrieval across languages. ArXiv **abs/2109.05125** (2021)
21. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q.V., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: ICML (2021)
22. Joulin, A., van der Maaten, L., Jabri, A., Vasilache, N.: Learning visual features from large weakly supervised data. In: ECCV (2016)
23. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Communications of the ACM **60**, 84 – 90 (2012)
24. Li, A., Jabri, A., Joulin, A., van der Maaten, L.: Learning visual n-grams from web data. 2017 IEEE International Conference on Computer Vision (ICCV) pp. 4193–4202 (2017)
25. Li, Y., Liang, F., Zhao, L., Cui, Y., Ouyang, W., Shao, J., Yu, F., Yan, J.: Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. ArXiv **abs/2110.05208** (2021)
26. van den Oord, A., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. ArXiv **abs/1807.03748** (2018)
27. Quattoni, A., Collins, M., Darrell, T.: Learning visual representations using images with captions. 2007 IEEE Conference on Computer Vision and Pattern Recognition pp. 1–8 (2007)
28. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: ICML (2021)
29. Radford, A., Narasimhan, K.: Improving language understanding by generative pre-training (2018)
30. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M.S., Berg, A.C., Fei-Fei, L.: Imagenet large scale visual recognition challenge. International Journal of Computer Vision **115**, 211–252 (2015)
31. Sariyildiz, M.B., Perez, J., Larlus, D.: Learning visual representations with caption annotations. In: ECCV (2020)
32. Sharma, P., Ding, N., Goodman, S., Soricut, R.: Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In: ACL (2018)
33. Thomee, B., Shamma, D.A., Friedland, G., Elizalde, B., Ni, K.S., Poland, D.N., Borth, D., Li, L.J.: Yfcc100m: the new data in multimedia research. Commun. ACM **59**, 64–73 (2016)
34. Tian, Y., Henaff, O.J., Oord, A.v.d.: Divide and contrast: Self-supervised learning from uncurated data. arXiv preprint arXiv:2105.08054 (2021)
35. Torralba, A., Efros, A.A.: Unbiased look at dataset bias. CVPR 2011 pp. 1521–1528 (2011)
36. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., J'egou, H.: Training data-efficient image transformers & distillation through attention. In: ICML (2021)
37. Wu, Z., Xiong, Y., Yu, S.X., Lin, D.: Unsupervised feature learning via non-parametric instance-level discrimination. ArXiv **abs/1805.01978** (2018)

38. Yuan, X., Lin, Z.L., Kuen, J., Zhang, J., Wang, Y., Maire, M., Kale, A., Faieta, B.: Multimodal contrastive training for visual representation learning. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 6991–7000 (2021)
39. Zhai, X., Puigcerver, J., Kolesnikov, A., Ruyssen, P., Riquelme, C., Lucic, M., Djolonga, J., Pinto, A.S., Neumann, M., Dosovitskiy, A., Beyer, L., Bachem, O., Tschannen, M., Michalski, M., Bousquet, O., Gelly, S., Houlsby, N.: A large-scale study of representation learning with the visual task adaptation benchmark. arXiv: Computer Vision and Pattern Recognition (2019)