Learning Visual Representations via Language-Guided Sampling

Mohamed El Banani Karan Desai Justin Johnson
University of Michigan
{mbanani,kdexd,justincj}@umich.edu

Abstract

Although an object may appear in numerous contexts, we often describe it in a limited number of ways. Language allows us to abstract away visual variation to represent and communicate concepts. Building on this intuition, we propose an alternative approach to visual representation learning: using language similarity to sample semantically similar image pairs for contrastive learning. Our approach diverges from image-based contrastive learning by sampling view pairs using language similarity instead of handcrafted augmentations or learned clusters. Our approach also differs from image-text contrastive learning by relying on pre-trained language models to guide the learning rather than directly minimizing a cross-modal loss. Through a series of experiments, we show that language-guided learning yields better features than image-based and image-text representation learning approaches.

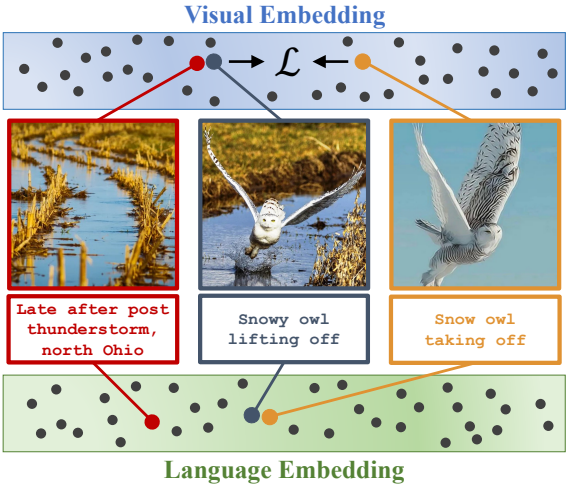

Figure 1. Language allows us to find conceptually similar image pairs even if they are visually dissimilar. We use those pairs for contrastive learning to learn generalizable visual features.

1. Introduction

Consider the images in Fig. 1, is the center image more similar to its left or right neighbor? Despite the difference in background and pose, it is clear that the right pair captures the same concept: a flying snow owl. Nevertheless, a self-supervised image model will judge the left pair as more similar. Human perception and language abstract away appearance differences to capture conceptual similarity rather than just visual similarity. Ideally, we could learn visual features that capture conceptual similarity and generalize effectively to other visual tasks. In this work, we show how language can be a proxy for conceptual similarity; allowing us to sample better pairs for contrastive learning and train more generalizable visual models.

Image-only contrastive learning uses visual similarity as a proxy for conceptual similarity. This is based on the observation that discriminative approaches can discover inter-class similarity–*e.g.*, cheetahs are similar to lions– without requiring explicit annotations [106]. The core idea is to train a discriminative model where each instance is treated as a separate class, and the model is trained to map augmented views of the same image to similar features [12–15, 106]. While successful, instance discrimination ignores the similarity between different instances as it assumes all other images are unrelated. Later work focused on inter-image relationships by estimating clusters [3, 9, 10] or finding nearest neighbors [28]. However, those relationships are estimated using visual embeddings; resulting in visually, rather than conceptually, similar pairs.

Language similarity is a strong proxy for semantic relationships. Consider the example in Fig. 1; images that depict the same concept are often described similarly. Radford *et al.* [76] propose language-image contrastive learning by mapping images and text to a shared representation space and achieve impressive generalization capabilities. However, it is unclear whether forcing models to map onto a shared space is optimal for visual learning. Although linguistic and visual similarity might align for similar instances, it is unclear whether all distances in one space should map exactly to the other. Instead of learning a joint vision-and-language representations, we argue that it is better to use linguistic similarity to guide visual learning.

To this end, we propose *language-guided contrastive learning*: a simple adaptation to contrastive learning that uses language models to find conceptually-similar image pairs for visual learning. Our approach is motivated by the observation that language models, despite never training on visual data, can still be used to sample caption pairs that belong to conceptually similar images, as seen in Fig. 2. Such sampled images exhibit desirable variations in pose, lightning, and context which are very different from hand-crafted augmentations which can be ill-suited to downstream tasks [108] or too focused on background textures [81]. We use the sampled pairs instead of image augmentations within standard self-supervised visual learning approaches such as SimCLR [12], SimSiam [15], and SLIP [67]. Our approach departs from image-only contrastive learning by relying on conceptually-similar image pairs rather than visually similar augmentations or cluster-assignment. We also depart from image-text pre-training by allowing the model to be guided by language similarity rather than learning a joint embedding space.

We conduct a series of controlled experiments to analyze our approach and compare it to commonly used representation learning paradigms on generalization to downstream classification tasks. In controlled settings, our approach outperforms all baselines on linear probe and few-shot classification on a range of downstream classification datasets. Our analysis suggests that while learning multi-modal joint embeddings can result in good representations, it is better to use one modality to guide the training of the other. Furthermore, we find that our approach is robust to the specific choice of sampling strategy or language model. Our code and pre-trained models are available at https://github.com/mbanani/lgssl.

2. Related Work

Visual Representation Learning aims to learn visual embedding spaces that capture semantics, with a typical focus on learning from scalable data sources. Broadly speaking, there are two general approaches: generative and discriminative. Generative approaches hypothesize that a model that can capture the image distribution will learn semantically relevant features [26, 31, 37, 70, 98, 115]. In contrast, discriminative approaches posit that differentiating between images will give rise to better features. This idea can be traced by to early work on metric learning [18] and dimensionality reduction [35], and is clearly seen for supervised classification models [84]. More recently, Wu *et al*. [106] proposed treating each image as a separate class and using augmented images as class instances to relieve the need for human annotation. This was followed by papers that simplified this approach [12–14, 38] and proposed non-contrastive variants [15, 34]. While those approaches have been successful, the utility of augmentation-based self-supervised learning has been questioned [68, 108] with follow-up work proposing the use of objectness [66, 75] and saliency [81] to alleviate some of those concerns. While we share the goal of visual representation learning, we question the reliance on image augmentations for training and propose using language models to learn for conceptually-similar images.

Language-supervised vision pre-training aims to learn visual representations from language data. Early work of Li *et al*. [57] trained n-gram models using YFCC [93] images and user-tag metadata. While some works learn joint vision-and-language representations for tasks like visual question answering [2, 33, 45, 118], visual reasoning [50, 89, 113], and retrieval [72, 112], we are interested in using language to learn better visual representations [23, 23, 76, 80, 88]. Early works used language modeling as a pretext task for visual learning [23, 80], but contrastive approaches quickly gained more popularity due to their relative simplicity and generalization capabilities [47, 76]. Follow-up work extended the contrastive formulation to learn dense features [109, 111] or used additional self-supervised losses to improve performance and data efficiency [21, 56, 59, 67]. While we share the motivation of using language for visual learning, we focus on learning visual representations by using linguistic guidance from pre-trained language models.

Leveraging structure in the data. This is commonly done in dense feature learning, where optical flow [36, 46, 82, 101] or 3D transformations [29, 44, 83, 90, 105] provide natural associations between image patches. For images, prior approaches used class names [51, 79], class hierarchies [58, 110], meta data [32, 48, 57] or clustering [3, 9, 10, 94, 117] to improve learning and inference. Within contrastive learning, clustering has been a popular choice for leveraging dataset structure. The intuition is that natural clusters emerge in feature spaces that can provide an additional training signal or useful pseudo-labels. While such approaches work well on curated datasets (*e.g*., ImageNet) where the label set provides an estimate of the number of clusters, it struggles with imbalanced and uncurated data [4]. Other approaches sample nearest neighbors as a feature-driven within-domain augmentation [28, 59]. While these approaches differ in how they extract inter-instance relationships, they all use within-domain feature similarity to sample positive pairs or clusters and hence do not leverage the rich cross-modal relationships. Closest to our work is Han *et al*. [36] who propose a co-training [6] scheme for jointly learning image and optical flow representations. We share their motivation of using similarity in one space (language) to learn in another (vision). Furthermore, instead of relying on co-training on the same dataset, we extract distances from a text-only language model, allowing us to leverage unaligned data.

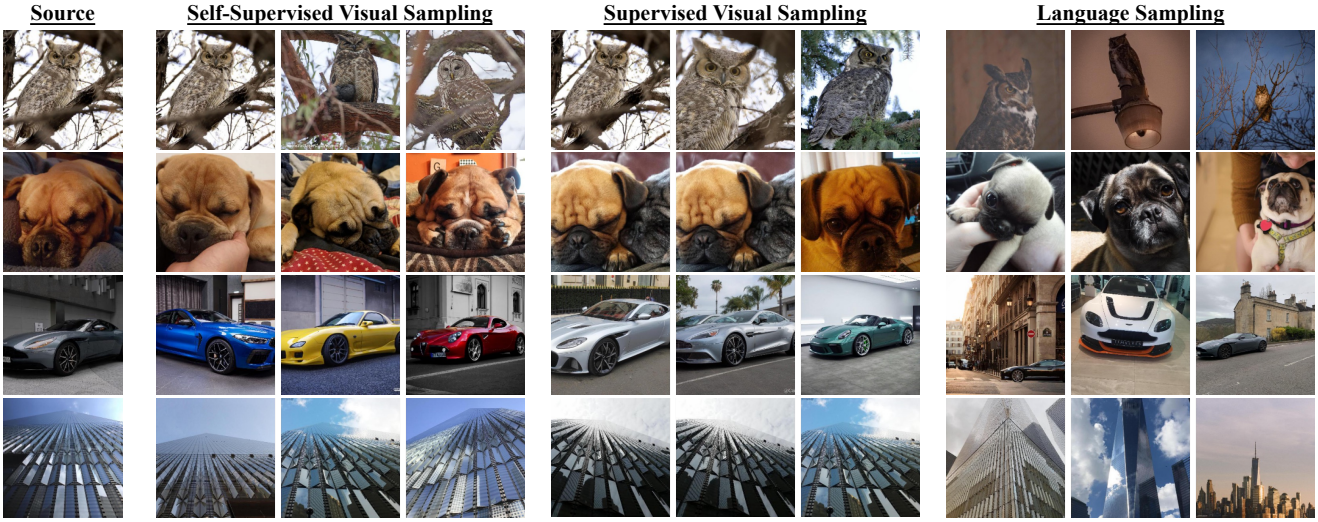

Figure 2. **Language sampling yields semantically-similar and visually-diverse image pairs.** We sample the three nearest neighbors using a self-supervised visual model [13], an ImageNet supervised model [27], and a self-supervised language model [78]. While visual sampling yields visually similar pairs, language sampling yields semantically relevant and visually diverse images. We argue that the combination of semantic consistency and visual diversity are better for learning generalizable features.

3. Method

The goal of this work is to learn visual representations that can generalize to other datasets. We extend image-only contrastive learning beyond hand-crafted augmentations and visually-sampled clusters to learn from conceptually similar images. Through learning to associate images that depict the same *visual concept*, models can learn visual invariances that more closely capture human semantics. To achieve this, we propose sampling image pairs that have similar captions using a pre-trained sentence encoder [78] and using them for contrastive learning. This work does not propose a new model or loss but rather a novel way of sampling image views that is applicable to a variety of approaches and losses for learning visual representations.

3.1. Learning from Conceptual Similarity

Instance discrimination has been the dominant task for visual representation learning. Its core intuition is that visual similarity is a good proxy for semantic similarity. The standard approach generates positive *view* pairs using image augmentations and maximizes their embedding similarity, with or without negative views. While there has been a large number of contrastive learning approaches, view pair generation has largely remained the same. Other methods use visual feature similarity to learn prototypes [3, 9, 10] or sample previously seen instances [28] for contrastive learning. While these approaches extend beyond instances and consider relations in the dataset, they still rely on visual similarity to generate their contrastive pairs. This limits the visual invariances that they can learn [108].

We propose training models to identify the same visual *concept* instead of the same *instance*. Our key observation is simple: images that have similar captions often depict similar concepts regardless of the actual appearance similarity. This can be clearly seen in Fig. 2. Nearest neighbors in visual representation space depict objects in similar scenes and poses, with self-supervised models showing some color invariances due to color augmentation. Conversely, similarly captioned images depict objects in different colors, poses, and contexts. This makes language-sampled images an excellent source for visual representation learning as they implicitly capture human-like visual invariances.

3.2. Sampling Image Pairs using Language

Given a captioned image dataset, we want to sample image pairs that have very similar captions. While caption similarity may be a good proxy for conceptual similarity, measuring caption similarity is a challenge on its own. Traditional metrics such as BLEU [71] and CIDER [96] rely on n-gram overlap, which can be too sensitive to phrasing and sentence structure. This makes them ill-suited for our needs. Other metrics such as SPICE [1] account for such variety by comparing parse trees; however, they still can not account for different wording choices. Inspired by advances in language models as well as approaches like BERTScore [116] and CLIPScore [43], we use a pre-trained sentence encoder to compute caption similarity.

Sentence encoders are trained to extract sentence-level features [52, 61, 78]. We use SBERT [78], which fine-tunes a pre-trained language model to allow it to better capture semantic similarity using feature cosine distance.

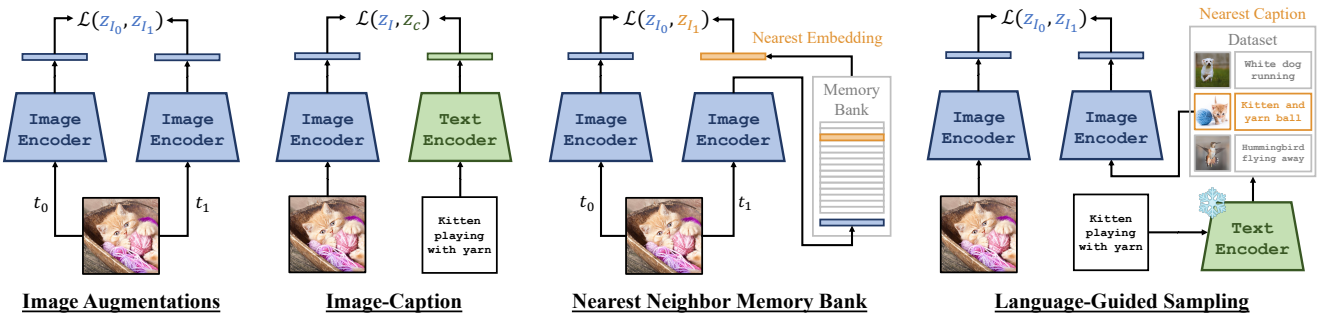

Figure 3. **Contrasting Contrastive Formulations**. While image-only and image-text contrastive learning directly extract views from the instance, nearest-neighbor methods rely on a memory bank of previously extracted features for training. In contrast, our approach samples nearest neighbors in caption embedding space using a pretrained language model and use the associated image for contrastive learning.

SBERT is trained in two stages: first, a language backbone is trained using a standard self-supervised task such as masked [25, 65] or permuted [87] language modeling; second, the language modeled is fine-tuned via contrastive learning on a large combined dataset of 1 billion sentence pairs. Fine-tuning the model in a contrastive way simplifies downstream usage as it allows features to be compared directly using cosine similarity. We use an SBERT [78] model with an MPNet [87] backbone. However, we find that our formulation is not sensitive to the choice of language encoder, as shown in Tab. 5c.

Finally, we sample the nearest neighbors for all captions in the language embedding space. We leverage modern similarity search libraries [49] to perform the nearest neighbor search quickly, despite the large dataset size. For example, nearest neighbor sampling runs in under 3 hours for RedCaps (12 million instances) on 4 GPUs, with 43 minutes spent on feature extraction and 117 minutes on nearest neighbor search. Furthermore, we find that we could further reduce the complexity of the sampling by only searching within subsets of the data as shown in Appendix E.

3.3. Language-Guided Visual Learning

Our approach is applicable to several representation learning methods as it only changes the training view pairs. We focus on contrastive learning since its fairly minimal setting allows us to analyze the impact of language guidance with minimal confounding factors. We train SimCLR with the language-sampled pairs and refer to it as LGSimCLR. We also evaluate the impact of language guidance on SimSiam [15] and SLIP [67], and find that they can similarly benefit from language guidance. We only use random cropping for image augmentations since language-sampled pairs are naturally augmented versions of each other and find that additional augmentations are not helpful. For LGSLIP, we match their setup by applying the CLIP loss only between the source's image and caption, ignoring an additional loss between the nearest neighbor image and its caption.

4. Experiments

Our experiments evaluate the efficacy of learning visual features from conceptually similar images. We hypothesize that a model trained with language guidance will learn useful visual invariances and better generalize to downstream tasks. We are interested in answering these questions: Does language guidance improve generalization over other pre-training approaches? Does language guidance generalize to other datasets and pre-training approaches? How can language be used for visual pre-training?

4.1. Experimental Setup

We formulate our experimental setup to compare the efficacy of different learning signals. We train models with language-guided sampling and compare them with image-only self-supervised models and image-text contrastive models. We are interested in conducting controlled experiments for a fair comparison.

Recent work in self-supervised learning has demonstrated the impressive impact of scaling [12, 76, 114]. While such work has shown impressive performance, it has complicated the evaluation as different models are trained on different pretext tasks on different datasets using varying amounts of compute and training recipes. Furthermore, replication is difficult, if not impossible, due to the unavailability of training data or prohibitive compute requirements. Fortunately, several papers report results that indicate that performance patterns often hold at smaller scales [10, 12, 21, 67, 76]. Hence, we conduct our experiments at a scale that allows us to perform a comprehensive evaluation and permits replication by others.

We conduct our experiments with a standard backbone [39] on publicly available datasets [11, 24, 85]. To account for variation in training recipes, we retrain all methods from scratch using the same training recipe. We scale down experiments to a level that permits fair comparisons and replication. We also provide system-level comparisons in Tab. 4 and scaling results in App. D.

Training details: We use a ResNet-50 backbone and train all models using the AdamW optimizer [63] with a learning rate of 10^{-3} and a weight decay of 10^{-2}. We use a cosine learning scheduler [62] with 5000 warm-up steps. Models are trained using a batch size of 512 for 250k steps; this corresponds to 10.5 epochs on RedCaps. We use a constant number of steps to permit meaningful comparisons between models trained on different datasets.

Evaluation setup: We evaluate all approaches using linear probe and fewshot classification on 15 classification datasets inspired by [53, 76]. We use the linear probe evaluation proposed by [53] and learn a single linear layer using logistic regression. We sweep over a range of cost values and choose the value with the best validation performance. We retrain a classifier on both train and validation splits and report test performance. We also evaluate all approaches on fewshot classification to understand their generalization ability. We use a weighted kNN classifier on frozen support features inspired by prior work showing its effectiveness for fewshot classification [102]. Please see Appendices A and B for more details on evaluation datasets and tasks.

Baselines: While there have been many proposed visual representation learning approaches, they can be grouped into several key directions that differ in the pretext task. We focus our comparison on a few representative approaches to explore the impact of the learning signal. We overview the baselines here and provide more details in Appendix C.

Many of our baselines are variants of contrastive learning as shown in Fig. 3. Contrastive approaches operate over paired source and target feature embeddings: $\boldsymbol{z}^s$ and $\boldsymbol{z}^t$. The goal is to maximize the similarity between the paired embeddings and minimize it with respect to all other embeddings. Given a batch size N and embedding dimension F, $\boldsymbol{z}^s, \boldsymbol{z}^t \in \mathbb{R}^{N \times F}$. The contrastive loss [86] is:

$$\mathcal{L}(z^s, z^t) = -\log \frac{\exp(\text{sim}(\boldsymbol{z}_i^s, \boldsymbol{z}_i^t)/\tau)}{\sum_{k=1}^{N} \exp(\text{sim}(\boldsymbol{z}_i^s, \boldsymbol{z}_k^t)/\tau)}, \quad (1)$$

where τ is a scaling parameter and $\text{sim}(\cdot, \cdot)$ is cosine similarity. Contrastive approaches primarily differ in how the embeddings are computed.

Image-Only Contrastive Learning contrasts features extracted from two randomly augmented *views* of the same image to perform instance discrimination [106]. We use SimCLR [12] as a representative approach due to its simplicity and strong performance.

Image-Text Contrastive Learning learns by contrasting features extracted from images and their captions. Unlike image-only approaches, this approach can learn semantics from the captions. Radford *et al*. [76] first proposed this approach and has had several follow-ups that augment it with additional self-supervised losses losses [56, 59, 67]. We use CLIP [76] and SLIP [67] due to their simplicity.

Nearest Neighbor Contrastive Learning contrast source embeddings with retrieved embeddings from a memory bank. The target features are used to retrieve the nearest neighbor embedding from a memory bank of previous batches. Dwibedi *et al*. [28] proposed this approach for image-only contrastive learning, while Li *et al*. [59] proposed adapting this loss for language embeddings. We use NNCLR [28] as Visual NNCLR and DeCLIP [59] with the CLIP and the language NNS losses as Language NNCLR.

Image-Only Non-Contrastive Learning deviates from the typical contrastive setup by learning without negative samples [15, 34]. We use SimSiam as a representative approach due to its simplicity and strong performance.

Cluster-based Contrastive Learning learn by contrasting image features with learned prototypes [3, 9, 10]. Prototypes are estimated via clustering or learned jointly with the feature encoder. Caron *et al*. [10] report that different cluster-based approaches perform similarly when provided with the same algorithmic advances. We use an adapted SwAV without the multi-crop augmentation strategy as it is equally applicable to other methods. We also compare against a pretrained SwAV checkpoint in Tab. 4.

4.2. Results

We train all approaches with a ResNet-50 backbone on RedCaps and report results in Tabs. 1 and 3. Our model outperforms all baselines with a significant margin for both evaluations. We analyze the results below through a series of questions.

Does language-guided sampling provide better training pairs than image augmentations? LGSimCLR greatly outperforms SimCLR despite using the same learning objective. By using language sampled pairs instead of image augmentations, LGSimCLR learns stronger invariances. We find that the largest gains arise in fine-grained datasets: Cars, CUB, and Food101. The performance gains can be explained by considering the critique of Xiao *et al*. [108]: the training augmentations dictate the invariances learned by SimCLR as shown in nearest neighbor samples in Fig. 2. Consider the third row of Fig. 2, while language sampling depicts three Aston Martin cars in different spots, visual nearest neighbors are sports cars in different poses and colors, closely resembling the flip and color augmentations used for training. Similarly in the first row of Fig. 2, visual nearest neighbors depict owls from different species in similar poses, while language sampling retrieves three great horned owls from different viewpoints. These trends are further amplified when features are used directly for fewshot classification. Language guidance allows us to capture relationships that go beyond visual similarity by training on image pairs that capture human semantics.

Table 1. **Linear Probe Evaluations.** We train ResNet-50 models on RedCaps and report performance of a linear probe using frozen features on 15 downstream tasks. Models are split based on whether or not they require caption images for training. LGSimCLR outperforms all previous approaches with strong performance gains for fine-grained classification datasets.

Model	Food-101	CIFAR-10	CIFAR-100	CUB	SUN397	Cars	Aircraft	DTD	Pets	Caltech-101	Flowers	STL-10	EuroSAT	RESISC45	PCAM	Mean
SwAV	63.6	81.3	57.5	21.6	47.5	22.9	35.4	68.1	61.1	70.5	78.0	87.7	94.3	79.9	84.3	63.6
SimSiam	64.1	79.9	56.1	28.2	48.3	29.5	41.2	66.2	69.1	73.6	83.6	85.7	94.4	82.1	83.3	65.7
Visual NNCLR	65.4	82.8	60.2	26.6	50.0	26.6	40.9	68.0	65.2	75.4	83.5	88.5	95.3	82.2	83.8	66.3
SimCLR	69.0	82.9	61.6	30.6	52.6	33.7	43.7	69.8	70.5	74.1	86.9	88.0	95.4	84.6	84.4	68.5
Language NNCLR	81.2	83.1	61.9	48.6	56.5	45.1	37.2	68.8	78.1	82.0	90.2	93.4	92.5	81.1	80.7	72.0
CLIP	80.9	84.7	62.7	50.4	57.4	45.8	36.7	67.6	79.8	84.0	91.0	93.5	93.9	82.2	82.6	72.9
SLIP	77.7	87.2	67.0	42.4	58.1	48.7	45.2	72.3	79.5	82.7	92.1	92.7	95.6	85.5	83.4	74.0
LGSimCLR (Ours)	83.2	87.8	69.0	59.3	60.3	62.3	53.4	71.2	81.8	89.4	95.9	94.0	95.6	88.0	81.1	78.2

Figure 4. **Nearest Neighbor methods are limited by the memory bank size.** Even with a large memory bank, the nearest embedding can still be unrelated to the source image while language sampling provides us with conceptually similar pairs.

Can we just sample nearest neighbors from previous batches? LGSimCLR outperforms NNCLR despite both relying on nearest neighbors. NNCLR uses the nearest feature embedding from a memory bank in the same modality. The quality of their retrieved samples is limited by the size of the memory bank. To demonstrate this, we visualize the nearest neighbors retrieved by NNCLR for different memory bank sizes in Fig. 4. We find that the retrieval quality is poor even for larger queues. Interestingly, we note that NNCLR also underperforms SimCLR on RedCaps, despite performing better on ImageNet. We posit that ImageNet's curated distribution explains this: a queue of 16k will most probably contain instances from each class, resulting in both visually and conceptually similar retrievals. Additionally, the quality of nearest neighbors is affected by the sampling feature space; features that are only trained on image augmentations will have limited invariances as shown in Fig. 2. We further explore the impact of sampling space on training in Sec. 4.3.

Table 2. **Language-guided contrastive learning outperforms image-text contrastive learning, regardless of text encoder.**

Objective	Text Encoder	Linear	Fewshot
Image-Text	Randomly-Initialized	72.9	77.5
	Frozen SBERT	71.8	77.1
Image-Image	Frozen CLIP (RedCaps)	78.3	82.4
	Frozen SBERT	78.2	82.5

Can cluster-based approaches learn better features? Similar to nearest-neighbor sampling, clustering is performed using visual similarity. Furthermore, it is based on an estimated number of clusters in the training dataset. Although this can be determined for ImageNet due to its known class structure, the number of clusters in an arbitrary uncurated dataset is unknown. This results in a large performance drop, as seen in Tab. 1 and Tab. 3. On the other hand, sampling related pairs assumes no global structure within the data and hence is able to better capture inter-instance similarity. This results in nearest-neighbor sampling outperforming clustering and both being outperformed by contrastive learning and language-guided contrastive learning.

Should we use language for guidance or supervision? Our experiments indicate that LGSimCLR outperforms both CLIP and SLIP. We consider two possible explanations: (a) SBERT extracts better language embeddings than CLIP can learn from the data, or (b) language-guided contrastive learning is a better training objective than image-text contrastive learning. To evaluate this, we compare four models in Tab. 2. The first two models use CLIP's training objective: the first model uses a randomly initialized language encoder, similar to CLIP. The second model uses a frozen SBERT model as the language encoder and only trains the projection layers. The second two models use LGSimCLR's training objective but sample pairs using a

Table 3. **Few-Shot Evaluations.** We train ResNet-50 models on RedCaps and report 5-way, 5-shot classification performance. We observe that language results in huge performance gains as shown by the performance of CLIP and LGSimCLR. Furthermore, the use of any augmentations hurts performance as seen by SLIP's drop in performance.

Model	Food-101	CIFAR-10	CIFAR-100	CUB	SUN397	Cars	Aircraft	DTD	Pets	Caltech-101	Flowers	STL-10	EuroSAT	RESISC45	Mean
SwAV	64.5	54.0	61.8	45.8	84.9	36.5	34.1	74.8	66.5	78.1	75.5	72.6	80.4	72.9	64.5
SimSiam	63.9	49.9	57.2	49.5	84.5	39.3	37.9	75.7	67.8	79.7	81.5	69.6	80.6	79.4	65.5
Visual NNCLR	65.6	54.1	61.7	45.8	85.3	37.9	34.9	75.2	67.3	81.1	75.4	74.3	83.6	76.7	65.6
SimCLR	66.9	45.7	51.0	51.5	87.1	44.0	38.4	77.6	70.1	80.0	86.9	69.6	83.5	81.3	66.7
Language NNCLR	89.3	65.3	73.4	78.6	90.8	68.4	40.4	75.2	78.8	90.9	94.3	89.6	75.2	71.9	77.3
CLIP	88.9	64.6	73.1	78.3	90.9	69.7	40.7	75.7	77.5	91.6	94.7	89.8	75.3	74.8	77.5
SLIP	81.5	63.5	70.8	63.1	91.3	62.9	42.1	79.6	76.4	88.4	92.2	83.4	82.7	80.8	75.6
LGSimCLR (Ours)	90.3	66.3	75.5	83.1	92.7	77.6	50.6	81.1	84.1	95.4	97.6	86.5	85.0	89.0	82.5

pre-trained language-only SBERT or the language encoder from a CLIP model trained on RedCaps. We find that image-image contrastive learning yields better visual features for both setups. While CLIP does not benefit from an SBERT backbone, LGSimCLR benefits from sampling using a language encoder trained on the same dataset. This suggests that learning joint embeddings results in worse visual features than language-guided learning.

System-level comparisons: We compare LGSimCLR with publicly-available checkpoints of prior approaches; see Appendix C for details. We emphasize that while the experiments reported in Tabs. 1 and 3 were done in a controlled setup (same batch size, training data, optimizer), the system level comparisons are trained on different datasets with different training recipes and enhancements to further boost performance; *e.g.*, large batch sizes, longer training, multi-crop augmentation. Furthermore, it has been shown that models trained on ImageNet implicitly benefit from its curated nature [4, 67]. Nevertheless, our approach still outperforms prior self-supervised approaches. We fall short of CLIP's ResNet-50 due to its training scale; $64\times$ larger batch, $32\times$ larger dataset, and $75\times$. We also observe that ImageNet-supervised ResNet-50 achieves better few-shot performance. Examining the performance breakdown in Tab. 10, we find the improvement mainly comes from CIFAR10, CIFAR100, and Pets. We posit that this can be explained by ImageNet's class structure: mostly pets with a large overlap with CIFAR's classes.

4.3. Analysis

We now analyze language-guided contrastive learning by evaluating the impact of pre-training data, the choice of embedding space, and the pretext task. By understanding the impact of those choices, we can better understand what the model is learning.

Table 4. **ResNet-50 System Level Comparisons.** We outperform prior self-supervised approaches despite them benefiting from ImageNet's curation for training and using larger batch sizes. CLIP outperforms us due to the scale of its training.

	Batch	# Img Updates	Dataset	Linear	Fewshot
Supervised [104]	1024	1.3×10^8	ImageNet	78.0	85.7
SimSiam [15]	512	1.3×10^8	ImageNet	72.9	78.7
SimCLR [13]	4096	1.0×10^9	ImageNet	75.4	77.4
MoCo [16]	4096	1.3×10^8	ImageNet	77.7	80.1
SwAV [10]	4096	1.3×10^8	ImageNet	78.2	78.5
CLIP [76]	32768	1.0×10^{10}	CLIP	81.8	87.8
LGSimCLR	512	1.3×10^8	RedCaps	78.2	82.5

Approach generality: We extend language guidance to other contrastive approaches: SimSiam and SLIP. We observe that language guidance uniformly improves performance for all methods, as shown in Tab. 5a. Furthermore, the difference between SimCLR and SLIP shrinks when adding language guidance. This suggests that language guidance provides the model with similar semantics to the ones learned from an image-text contrastive loss, resulting in diminished gains from the additional image-text loss.

Impact of training dataset: We train our model on four datasets: CC3M [85], CC12M [11], RedCaps-2020, and RedCaps [24]. In Tab. 5b, we observe that larger datasets result in stronger performance, indicating that our approach could scale well with even larger datasets. Furthermore, we observe that RedCaps results in better performance than Conceptual Captions. This may be attributed to the higher quality of captions in RedCaps; while the alt-text captions CC3M and CC12M can be short and contain image metadata, RedCaps captions are diverse, longer, and more descriptive. This allows our model to sample more interesting visual pairs that capture more visual diversity. We provide qualitative results in Appendix F to support this.

	Image Aug.		Language	
	Linear	Fewshot	Linear	Fewshot
SimSiam	65.7	65.5	71.2	75.7
SimCLR	68.5	66.7	78.2	82.5
SLIP	74.0	75.6	78.8	82.8

(a) **Approach Generality**

	Size	Linear	Fewshot
CC3M	2.7M	71.5	76.3
CC12M	10.9M	76.8	81.9
RedCaps 2020	3.2M	73.8	78.8
RedCaps	12.0M	78.2	82.5

(b) **Impact of training dataset**

	Linear	Fewshot
SBERT (MPNet)	78.2	82.5
SBERT (MiniLM)	78.6	83.3
CLIP Language (ViT-B/32)	78.3	83.1
FastText BoW	76.1	80.9
ImageNet-supervised	78.3	81.8
SimCLR (ImageNet)	73.1	74.6

(c) **Impact of sampling space**

Table 5. **Analysis Experiments.** We conduct a series of analysis experiments to understand language-guided contrastive learning. The results indicate that language sampling is beneficial to several formulations and scales well with larger datasets. Furthermore, while language sampling consistently results in good pairs for training, visual sampling only helps if it has access to semantics through supervision.

Impact of sampling space: The idea of using offline nearest-neighbor sampling does not require a specific language model or even a specific modality. We explore other choices for embedding space: four sentence encoders and two image models. In our experiments, we use SBERT's MPNet model [78, 87]; the highest performing SBERT model for sentence similarity. We compare it to two other sentence transformers: a smaller SBERT model, MiniLM [100], and the language encoder from CLIP [76]. We also compared against a bag-of-words (BoW) sentence encoder that uses FastText [7] embeddings. Results are in Tab. 5c. While we expected that using CLIP for sampling would improve performance due to its multimodal training, we were surprised that MiniLM also improved performance despite its lower performance on language tasks. We find that pairs obtained using a BoW model result in a weaker performance which might hint at the importance of contextual sentence embeddings. Nevertheless, the BoW-sampled pairs still result in higher performance than all the other baselines on RedCaps.

We also consider training with pairs sampled using two visual models: ImageNet-supervised ResNet-50 [104] and ImageNet-trained SimCLR [13]. We find that using a visual model for sampling is only beneficial if the visual model captures semantic relations; *e.g.*, through supervised training. Using a self-supervised language model results in a strong drop in performance relative to the other sampling spaces. Nevertheless, it still allows the model to achieve better performance than using a self-supervised visual approach on the same data. This indicates that while language is a better modality to use, "sample-guided" contrastive learning can still achieve a stronger performance than only using self-supervised learning.

Limitations: We observe a few limitations in our approach. Image captions can be noisy, vague, and often omit obvious relations in the image [5]. While this broadly affects image-language models, it can result in us retrieving unrelated image pairs. For example, captions like "*I found this in the garden*" or "*Photo from our family trip*" could describe a large range of images, some of which are unrelated. We expand on this in Appendix F. Image descriptions also depend on the context and the perceiver; *e.g.*, a tourist and an art curator will describe artwork in very different ways. We observe that descriptions in topic-focused subreddits (*e.g.*, **`r/birdpics`** and **`r/woodworking`**) are more specific than in generic subreddits (*e.g.*,**`r/itookapicture`** and **`r/pics`**). Our experiments in Appendix E support this observation. Since a caption only captures one aspect of the image, sampled pairs can be similar for a variety of reasons. Allowing the model to condition the feature extraction or similarity calculation on captions could alleviate this issue.

5. Conclusion

We propose using language to find conceptually similar images for contrastive learning. This is based on a simple observation: people describe an object in similar ways even when it appears in different contexts. We use pre-trained language models to sample similar captions and use the captioned images for contrastive learning. We hypothesize that using language guidance instead of image augmentations would result in learning more human-like invariances.

We evaluate our approach on multiple train and test datasets and find that it outperforms previous self-supervised and image-text contrastive models. Our analysis demonstrates the utility of using nearest-neighbor instances for training and the superiority of language sampling over other approaches for unlabeled datasets. Our findings align with prior work that critiques the use of image augmentations [81, 108] and shows the utility of cross-modal guidance [36] and intra-instance relationships [28, 51]. Our results demonstrate the potential of incorporating language guidance in contrastive learning. We hope that future work will explore scaling up our approach to larger and more diverse datasets, as well as modeling approaches that further integrate language into the learning process.

Acknowledgments: We thank Richard Higgins, Ashkan Kazemi, and Santiago Castro for many helpful discussions, as well as David Fouhey, Ziyang Chen, Chenhao Zheng, Fahad Kamran, and Dandan Shan for their feedback on early drafts. This project was funded under the Ford-UM Alliance partnership; we thank Alireza Rahimpour, Devesh Upadhyay, and Ali Hassani from Ford Research for their support and discussion.

References

[1] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. SPICE: Semantic propositional image caption evaluation. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2016. 3

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. VQA: Visual question answering. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 2015. 2

[3] Yuki Markus Asano, Christian Rupprecht, and Andrea Vedaldi. Self-labelling via simultaneous clustering and representation learning. In *International Conference on Learning Representations*, 2019. 1, 2, 3, 5

[4] Mido Assran, Randall Balestriero, Quentin Duval, Florian Bordes, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, and Nicolas Ballas. The hidden uniform cluster prior in self-supervised learning. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023. 2, 7

[5] Hessam Bagherinezhad, Hannaneh Hajishirzi, Yejin Choi, and Ali Farhadi. Are elephants bigger than butterflies? reasoning about sizes of objects. In *AAAI Conference on Artificial Intelligence*, 2016. 8

[6] Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of Annual Conference on Computational learning theory*, 1998. 2

[7] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*, 2016. 8

[8] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *European Conference on Computer Vision*, 2014. 15

[9] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European conference on computer vision (ECCV)*, pages 132–149, 2018. 1, 2, 3, 5

[10] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 2020. 1, 2, 3, 4, 5, 7, 15, 21, 22, 23

[11] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12M: Pushing Web-Scale Image-Text Pre-Training To Recognize Long-Tail Visual Concepts. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 4, 7

[12] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A Simple Framework for Contrastive Learning of Visual Representations. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2020. 1, 2, 4, 5, 15, 16

[13] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. Big self-supervised models are strong semi-supervised learners. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 2020. 1, 2, 3, 7, 8, 15, 16, 20, 21, 22, 23

[14] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 1, 2

[15] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1, 2, 4, 5, 7, 21, 22, 23

[16] Xinlei Chen*, Saining Xie*, and Kaiming He. An empirical study of training self-supervised vision transformers. *arXiv preprint arXiv:2104.02057*, 2021. 7, 15, 21, 22, 23

[17] Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883, Oct 2017. 15

[18] Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005. 2

[19] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, and A. Vedaldi. Describing textures in the wild. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 15

[20] Adam Coates, Andrew Ng, and Honglak Lee. An Analysis of Single Layer Networks in Unsupervised Feature Learning. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2011. https://cs.stanford.edu/~acoates/papers/coatesleeng_aistats_2011.pdf. 15

[21] Yufeng Cui, Lichen Zhao, Feng Liang, Yangguang Li, and Jing Shao. Democratizing contrastive language-image pre-training: A CLIP benchmark of data, model, and supervision. In *ICML Workshop on Pre-training: Perspectives, Pitfalls, and Paths Forward*, 2022. 2, 4

[22] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009. 15, 18

[23] Karan Desai and Justin Johnson. Virtex: Learning visual representations from textual annotations. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11162–11173, 2021. 2

[24] Karan Desai, Gaurav Kaul, Zubin Aysola, and Justin Johnson. RedCaps: Web-curated image-text data created by the people, for the people. In *NeurIPS Datasets and Benchmarks*, 2021. 4, 7

[25] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics – Human Language Technologies (NAACL HLT)*, 2018. 4

[26] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 2015. 2

[27] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021. 3, 15, 20

[28] Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. With a little help from my friends: Nearest-neighbor contrastive learning of visual representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9588–9597, October 2021. 1, 2, 3, 5, 8, 16

[29] Mohamed El Banani, Luya Gao, and Justin Johnson. Unsupervisedr&r: Unsupervised point cloud registration via differentiable rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7129–7139, 2021. 2

[30] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *Computer Vision and Pattern Recognition Workshop*, 2004. 14, 15

[31] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018. 2

[32] Yunchao Gong, Qifa Ke, Michael Isard, and Svetlana Lazebnik. A multi-view embedding space for modeling internet images, tags, and their semantics. In *IJCV*, 2014. 2

[33] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[34] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 2, 5

[35] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2006. 2

[36] Tengda Han, Weidi Xie, and Andrew Zisserman. Self-supervised co-training for video representation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 2, 8

[37] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[38] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[39] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 4, 15

[40] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2019. 15

[41] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 2021. 15, 18

[42] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 15, 18

[43] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021. 3

[44] Ji Hou, Saining Xie, Benjamin Graham, Angela Dai, and Matthias Nießner. Pri3d: Can 3d priors help 2d representation learning? In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 2021. 2

[45] Drew A Hudson and Christopher D Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[46] Allan Jabri, Andrew Owens, and Alexei Efros. Space-time correspondence as a contrastive random walk. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 2

[47] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2021. 2

[48] Justin Johnson, Lamberto Ballan, and Li Fei-Fei. Love thy neighbors: Image annotation by exploiting image metadata. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 2015. 2

[49] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 2019. 4

[50] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014. 2

[51] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning.

In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 2, 8

[52] Ryan Kiros, Yukun Zhu, Russ R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Skip-thought vectors. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2015. 3

[53] Simon Kornblith, Jonathon Shlens, and Quoc V Le. Do better imagenet models transfer better? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2661–2671, 2019. 5, 14

[54] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13)*, Sydney, Australia, 2013. 15

[55] Alex Krizhevsky. Learning multiple layers of features from tiny images, 2009. 15

[56] Janghyeon Lee, Jongsuk Kim, Hyounguk Shon, Bumsoo Kim, Seung Hwan Kim, Honglak Lee, and Junmo Kim. Uniclip: Unified framework for contrastive language-image pre-training. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 2, 5

[57] Ang Li, Allan Jabri, Armand Joulin, and Laurens van der Maaten. Learning visual n-grams from web data. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 2017. 2

[58] Aoxue Li, Tiange Luo, Zhiwu Lu, Tao Xiang, and Liwei Wang. Large-scale few-shot learning: Knowledge transfer with class hierarchy. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[59] 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. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022. 2, 5, 16

[60] Dong C Liu and Jorge Nocedal. On the limited memory bfgs method for large scale optimization. *Mathematical programming*, 1989. 14

[61] Lajanugen Logeswaran and Honglak Lee. An efficient framework for learning sentence representations. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018. 3

[62] Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 5

[63] Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019. 5

[64] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 15

[65] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2013. 4

[66] Shlok Mishra, Anshul Shah, Ankan Bansal, Abhyuday Jagannatha, Abhishek Sharma, David Jacobs, and Dilip Krishnan. Object-aware cropping for self-supervised learning. *arXiv preprint arXiv:2112.00319*, 2021. 2

[67] Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. Slip: Self-supervision meets language-image pre-training. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2022. 2, 4, 5, 7, 15, 16

[68] Alejandro Newell and Jia Deng. How useful is self-supervised pretraining for visual tasks? In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[69] M-E. Nilsback and A. Zisserman. Automated flower classification over a large number of classes. In *Proceedings of the Indian Conference on Computer Vision, Graphics and Image Processing*, Dec 2008. 15

[70] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 2

[71] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the Annual Meeting on Association for Computational Linguistics (ACL)*, 2002. 3

[72] Yookoon Park, Mahmoud Azab, Bo Xiong, Seungwhan Moon, Florian Metze, Gourab Kundu, and Kirmani Ahmed. Normalized contrastive learning for text-video retrieval. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2022. 2

[73] O. M. Parkhi, A. Vedaldi, A. Zisserman, and C. V. Jawahar. Cats and dogs. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. 15

[74] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 14

[75] Xiangyu Peng, Kai Wang, Zheng Zhu, Mang Wang, and Yang You. Crafting better contrastive views for siamese representation learning. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16031–16040, 2022. 2

[76] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2021. 1, 2, 4, 5, 7, 8, 14, 16, 21, 22, 23

[77] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *Proceedings of the International Conference on Machine Learning (ICML)*, 2019. 15, 18

[78] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural*

Language Processing. Association for Computational Linguistics, 11 2019. 3, 4, 8, 14, 20

[79] Karsten Roth, Oriol Vinyals, and Zeynep Akata. Integrating Language Guidance into Vision-based Deep Metric Learning. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[80] Mert Bulent Sariyildiz, Julien Perez, and Diane Larlus. Learning visual representations with caption annotations. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2020. 2

[81] Ramprasaath R Selvaraju, Karan Desai, Justin Johnson, and Nikhil Naik. Casting your model: Learning to localize improves self-supervised representations. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11058–11067, 2021. 2, 8

[82] Dandan Shan, Richard Ely Locke Higgins, and David Fouhey. Cohesiv: Contrastive object and hand embedding segmentation in video. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 2

[83] Jinghuan Shang, Srijan Das, and Michael S Ryoo. Learning viewpoint-agnostic visual representations by recovering tokens in 3d space. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 2

[84] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. In *CVPRW*, 2014. 2

[85] 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 the Annual Meeting of the Association for Computational Linguistics*, 2018. 4, 7

[86] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016. 5

[87] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. MPNet: Masked and Permuted Pre-training for language understanding. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 4, 8, 14

[88] Jonathan C Stroud, Zhichao Lu, Chen Sun, Jia Deng, Rahul Sukthankar, Cordelia Schmid, and David A Ross. Learning video representations from textual web supervision. *arXiv preprint arXiv:2007.14937*, 2020. 2

[89] Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs. In *Proceedings of the Annual Meeting on Association for Computational Linguistics (ACL)*, 2019. 2

[90] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. LoFTR: Detector-free local feature matching with transformers. *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[91] Igor Susmelj, Matthias Heller, Philipp Wirth, Prescott Jeremey, and Malte Ebner. Lightly. *GitHub. Note: https://github.com/lightly-ai/lightly*, 2020. 16

[92] TensorFlow Datasets, a collection of ready-to-use datasets. https://www.tensorflow.org/datasets. 14

[93] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. YFCC100M: The new data in multimedia research. *Communications of the ACM*, 2016. 2

[94] Yonglong Tian, Olivier J Henaff, and Aäron van den Oord. Divide and contrast: Self-supervised learning from uncurated data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. 2

[95] Yonglong Tian, Yue Wang, Dilip Krishnan, Joshua B Tenenbaum, and Phillip Isola. Rethinking few-shot image classification: a good embedding is all you need? In *Proceedings of European Conference on Computer Vision (ECCV)*, 2020. 14

[96] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. CIDEr: Consensus-based image description evaluation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 3

[97] Bastiaan S Veeling, Jasper Linmans, Jim Winkens, Taco Cohen, and Max Welling. Rotation equivariant cnns for digital pathology. In *International Conference on Medical image computing and computer-assisted intervention*, 2018. 15

[98] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2008. 2

[99] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 15, 18

[100] Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 8, 14

[101] Xiaolong Wang, Allan Jabri, and Alexei A Efros. Learning correspondence from the cycle-consistency of time. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[102] Yan Wang, Wei-Lun Chao, Kilian Q. Weinberger, and Laurens van der Maaten. Simpleshot: Revisiting nearest-neighbor classification for few-shot learning. *arXiv preprint arXiv:1911.04623*, 2019. 5, 14

[103] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology, 2010. 15

[104] Ross Wightman, Hugo Touvron, and Herve Jegou. Resnet strikes back: An improved training procedure in timm. In *NeurIPS Workshop on ImageNet: Past, Present, and Future*, 2021. 7, 8, 15, 21, 22, 23

[105] Xiaoshi Wu, Hadar Averbuch-Elor, Jin Sun, and Noah Snavely. Towers of babel: Combining images, language, and 3d geometry for learning multimodal vision. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 2021. 2

[106] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3733–3742, 2018. 1, 2, 5

[107] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba. SUN database: Large-scale scene recognition from abbey to zoo. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010. 15

[108] Tete Xiao, Xiaolong Wang, Alexei A Efros, and Trevor Darrell. What should not be contrastive in contrastive learning. In *International Conference on Learning Representations*, 2020. 2, 3, 5, 8

[109] Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. Groupvit: Semantic segmentation emerges from text supervision. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[110] Zhicheng Yan, Hao Zhang, Robinson Piramuthu, Vignesh Jagadeesh, Dennis DeCoste, Wei Di, and Yizhou Yu. Hd-cnn: hierarchical deep convolutional neural networks for large scale visual recognition. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 2015. 2

[111] 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. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022. 2

[112] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL*, 2014. 2

[113] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[114] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 4

[115] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2016. 2

[116] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019. 3

[117] Mingkai Zheng, Fei Wang, Shan You, Chen Qian, Changshui Zhang, Xiaogang Wang, and Chang Xu. Weakly supervised contrastive learning. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 2021. 2

[118] Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. Visual7w: Grounded question answering in images. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2