Learning Vision from Models Rivals Learning Vision from Data

Yonglong Tian[1,†] Lijie Fan[2,†,*] Kaifeng Chen[1] Dina Katabi[2] Dilip Krishnan[1] Phillip Isola[2]

[1]Google Research, [2]MIT CSAIL, [†]equal contribution

Github Repo: https://github.com/google-research/syn-rep-learn

Abstract

We introduce SynCLR, a novel approach for learning visual representations exclusively from synthetic images and synthetic captions, without any real data. We synthesize a large dataset of image captions using LLMs, then use an off-the-shelf text-to-image model to generate multiple images corresponding to each synthetic caption. We perform visual representation learning on these synthetic images via contrastive learning, treating images sharing the same caption as positive pairs. The resulting representations transfer well to many downstream tasks, competing favorably with other general-purpose visual representation learners such as CLIP and DINO v2 in image classification tasks. Furthermore, in dense prediction tasks such as semantic segmentation, SynCLR outperforms previous self-supervised methods by a significant margin, e.g.*, improving over MAE and iBOT by 6.2 and 4.3 mIoU on ADE20k for ViT-B/16.*

1. Introduction

Representation learning extracts and organizes information from raw, often unlabeled data. The quality, quantity, and diversity of the data determines how good a representation the model can learn. The model becomes a reflection of the collective intelligence that exists in the data. We get what we feed in.

Unsurprisingly, the current best-performing visual representation learning methods [59, 62] rely on large scale real datasets. However, the collection of real data has its own dilemmas. Collecting *large scale uncurated* data [71] is relatively cheap and thus quite achievable. However, for self-supervised representation learning, this approach exhibits poor scaling behavior –i.e., adding more uncurated data has little effect at large data scales [33, 80]. Collecting *small scale curated* data [21] also is achievable, but models trained in this way are limited to relatively narrow tasks. The ideal would be large scale curated datasets of real images, and

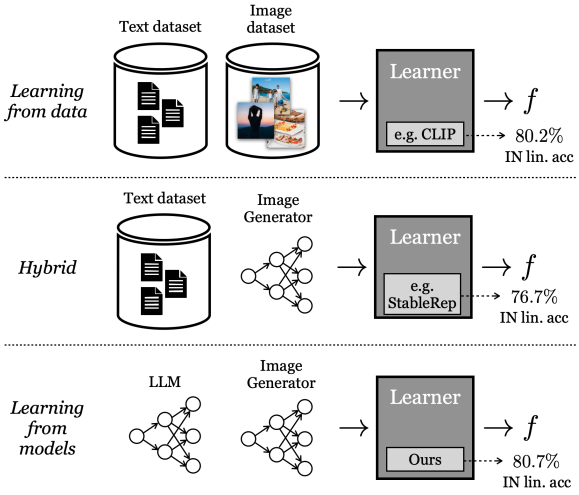

Figure 1. Three paradigms for visual representation learning. Top row: Traditional methods, such as CLIP [62], learn only from real data; Middle row: Recent methods, such as StableRep [81], learn from real text and generated images; Bottom row: Our method, SynCLR, learns from synthetic text and synthetic images, and rival the linear transfer performance of CLIP on ImageNet despite not directly observing any real data.

recent work has indeed shown that this can lead to strong performance gains at scale [59], but this path is costly to pursue.

To alleviate the cost, in this paper we ask if *synthetic data*, sampled from off-the-shelf generative models, is a viable path toward large scale curated datasets that can train state-of-the-art visual representations.

We call such a paradigm *learning from models*, in contrast to directly *learning from data*. Models have several advantages as a data source for building large scale training sets: via their latent variables, conditioning variables, and hyperparameters, they provide new controls for curating data; we will make use of these controls in the method we propose. Models also can be easier to share and store (because models are more compressed than data), and can produce an unlimited number of data samples (albeit with

*Work done while interning at Google.

finite diversity). A growing literature has studied these properties and other advantages (and disadvantages) of using generative models as a data source for training downstream models [3, 26, 40, 41, 69, 81]. Some of these methods use a *hybrid* mode – either mixing real and synthetic datasets [3] or needing a real dataset to generate another synthetic dataset [81]. Other methods try to learn representations from purely synthetic data [69] but lag far behind the best performing models. Instead, we show that *learning from models*, without training on any real data, can yield representations that match the top-performing representations learnt from real data. For instance, as illustrated in Figure 1, representations learnt by our method are able to transfer as well as OpenAI's CLIP [62] on ImageNet (both methods using ViT-B [24]).

Our approach leverages generative models to re-define the granularity of visual classes. As shown in Figure 2, consider we have four images generated using two prompts: "*a golden retriever, wearing sunglasses and a beach hat, rides a bike*" and "*a cute golden retriever sits in a house made of sushi*". Traditional self-supervised method such as SimCLR [13] will treat each of these images as a different class; embeddings for different images are pushed apart with no explicit consideration of the shared semantics between images. On the other extreme, supervised learning methods (*i.e*. SupCE) will regard all these images as a single class (e.g., "golden retriever"). This ignores nuances in the semantics of the images, such as the fact that the dogs are riding a bike in one pair of images and sitting inside a sushi house in the other pair of images. Instead, our method, SynCLR, treats *captions* as classes, *i.e*., each caption describes a visual class (this level of granularity was also explored in StableRep [81]). This allows us to group images by the concepts of "riding a bike" and "sitting in a sushi house", in addition to grouping by a coarser class label like "golden retrieval". This level of granularity is difficult to mine in real data, since collecting multiple images described by a given caption is non-trivial, especially when scaling up the number of captions. However, text-to-image diffusion models are fundamentally built with this ability; simply by conditioning on the same caption and using different noise inputs, a text-to-image diffusion model will produce different images that all match the same caption. In our experiments, we find the caption-level granularity outperforms both SimCLR and supervised training. Another advantage is that this definition of visual classes has good scalability. Unlike ImageNet-1k/21k where a given number of classes is fixed, we can augment existing classes (or data) in an online fashion, and theoretically scale up to as many classes as needed.

Our system consists of three steps. The first step is to synthesize a large corpus of image captions. We design a scalable approach by leveraging the in-context learning capability of large language models (LLMs), where we present

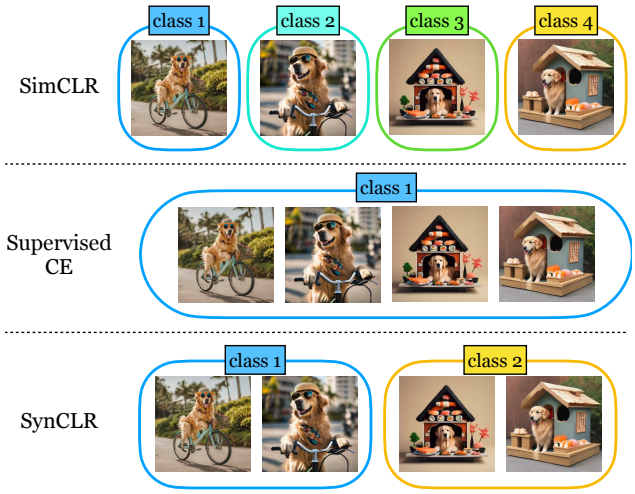

Figure 2. Different learning objectives treat classification granularity differently. These images are generated by two prompts "*a golden retriever, wearing sunglasses and a beach hat, rides a bike*" and "*a cute golden retriever sits in a house made of sushi*". SimCLR treats each image as a class, while supervised cross-entropy treats them all as the same "golden retrieval" class. The former does not consider shared semantics between images, and the latter is coarse-grained and ignores actions or relationships between subjects/background. Our approach, SynCLR, defines visual classes by sentences.

examples of word-to-caption translations. Next, a text-to-image diffusion model is adopted to synthesize multiple images for each synthetic caption. This yields a synthetic dataset of 600M images. Then we train visual representation models by a combination of multi-positive contrastive learning [43] and masked image modeling [98].

Our learned representations transfer well. With SynCLR pre-training, our ViT-B and ViT-L models achieve 80.7% and 83.0% top-1 linear probing accuracy on ImageNet-1K, respectively, which is on par with OpenAI's CLIP [62]. On fine-grained classification tasks, SynCLR outperforms CLIP by 3.3% for ViT-B and 1.5% for ViT-L, and performs similarly to DINO v2 [59] models, which are distilled from a pre-trained ViT-g model. For semantic segmentation on ADE20k, SynCLR outperforms MAE pre-trained on ImageNet by 6.2 and 4.1 in mIoU for ViT-B and ViT-L under the same setup, showing strong transfer ability for dense prediction tasks similar to DINO v2, which additionally involves a training period on 518x518 resolution images that SynCLR does not have.

2. Related Works

Self-supervised representation learning approaches in vision develop domain-specific pre-text tasks, such as colorization [94], rotation prediction [31], and solving jigsaw puzzles [56]. Domain-agnostic approaches have been pop-

ular, such as contrastive learning [6, 13, 35, 38, 57, 78, 87] and masked image modeling [2, 4, 5, 29, 39, 86, 90, 98]. Contrastive learning promotes invariance [79] for two views of the same image and pushes apart representations for different images [85] (or only invariance [11, 34]); the resulting representations yield strong performance for linear or zero-shot transfer. Masked image modeling reconstructs the pixels [39, 90] or local features [4], often producing excellent fine-tuning transfer performance, especially in dense prediction tasks [39]. The state of the art DINO v2 [59] leverages both approaches, and our approach shares a similar spirit.

Supervised learning [36, 45, 75] used to be the dominant approach for learning transferable visual representations for various tasks [23, 32, 72]. Recent studies [37, 49] has shown that, the transferability of representations learned in this way is limited, *e.g.*, pre-training has no improvement over random initialization for dense prediction tasks (*e.g.*, object detection) when the fine-tuning is long enough. Such limitation continues when the model has been scaled up to 22B [20]. An alternative paradigm learns visual representations from text supervision [42, 62], *e.g.*, CLIP [62]. This approach is more flexible (*i.e.*, not requiring classes) and provides richer supervision, often learning generalizable representations.

Generative models as representation learners. A number of papers have explored the representations that are learned by generative models for various recognition tasks [22, 48]. As might be expected intuitively, such models indeed learn especially good representations for dense tasks, such as optical flow estimation [70], semantic segmentation [8, 91], and depth estimation [95]. Another line of work [18, 47] adapt pre-trained diffusion models for zero-shot image recognition via analysis-by-synthesis. These approaches may need to be adapted when the architectures of the generative models change or a new family of generative model emerge. Our approach treats images as universal interfaces with the hope of better generality.

Learning from synthetic data from generative models. Synthetic data has been explored to train machine learning models in various domains [27, 46, 53, 54, 65, 66, 74, 77, 92]. In computer vision, the utilization of synthetic data for training models is common, ranging from optical flow [52] and autonomous driving [1] to semantic segmentation [15] and human pose estimation [84]. Others [41, 50] have explored synthetic data for representation learning, with the predominant approach of altering the latent variables of deep generative models. Our approach aligns with this research paradigm, but it diverges in its use of text-to-image models, which have also been investigated by other researchers [40, 69, 99]. But they use synthetic data for supervised learning [26, 69]. The closet work is StableRep [81], which also conducts representation learning but still needs a real text dataset.

3. Approach

In this paper, we study the problem of learning a visual encoder f in the absence of real images or textual data. Our approach hinges on the utilization of three key resources: a language generation model (g_1), a text-to-image generative model (g_2), and a curated list of visual concepts (C). Our exploration include three steps: (1) we employ g_1 to synthesize a comprehensive set of image descriptions T, which encompass the range of visual concepts in C; (2) for each caption in T, we generate multiple images using g_2, culminating in an extensive synthetic image dataset X; (3) we train on X to obtain a visual representation encoder f.

We use Llama-2 7B [83] and Stable Diffusion 1.5 [64] as g_1 and g_2, respectively, because of their fast inference speed. We anticipate that better g_1 and g_2 in the future will further enhance the effectiveness of this approach.

3.1. Synthesizing captions

To harness the capability of powerful text-to-image models for generating a substantial dataset of training images, we initially require a collection of captions that not only precisely depict an image but also exhibit diversity to encompass a broad spectrum of visual concepts.

We have developed a scalable approach to create such a large collection of captions, leveraging the in-context learning capability of LLMs [9]. Our method involves crafting specific prompt engineering templates that guide the LLM to produce the required captions. We start by gathering the concept list C from some existing datasets, such as ImageNet-21k [21] and Places-365 [96]. For each concept $c \in C$, we consider three straightforward templates to generate captions effectively.

- c –> *caption*. As the most direct and simple approach, we have the Llama-2 model sample a sentence for the concept c.
- c, bg –> *caption*. We combine the visual concept c with a background or setting bg. A naïve approach would randomly select both c and bg, where bg may correspond to a class name from a places dataset like [96]. However, this method often leads to unlikely combinations in the real world, such as a blue whale in a football field. Our ablation experiments demonstrate that this strategy results in suboptimal performance, likely because the generated captions fall far outside the training distribution of g_2. Instead, we employ GPT-4 [58] to generate a list of suitable backgrounds for the chosen concepts. This approach increases the likelihood of generating more plausible combinations, such as a tiger in a forest or a cat in a kitchen, enhancing the overall quality of the results.
- c, rel –> *caption*. Given a visual concept c, we consider pairing it with a positional relationship word, rel. Take for instance, if c signifies *cat* and rel translates to *in front*

Templates	In context examples
c –> *caption*	revolver –> Multiple antique revolvers lie on a wooden table, gleaming under soft, ambient light. closet –> The compact closet, brimming with clothes and shoes, exudes a feeling of organization. zebra –> A zebra is gallantly trotting across the vast, sunlit plains of the African savannah, creating a captivating black and white spectacle. bus station –> The bustling bus station thrums with restless energy, as travelers navigate through the crowded space, awaiting their journeys amid the echoes of departing buses.
c,bg –> *caption*	tiger, forest –> Two tigers are running together in the forest. lighter, motorhome –> In the cozy, cluttered environment of a well-traveled motorhome, a sleek silver lighter holds dominion on the rustic wooden table. sunset, lake –> Golden sunset hues reflect on a calm lake, silhouetting a lone canoeist against a backdrop of fiery clouds.
c,rel –> *caption*	kit fox, in front of –> A group of small, fluffy, golden kit foxes is playfully gathered in front of a lush, green, towering forest backdrop. cabbage, besides –> A vibrant image portrays a lush, green cabbage, glistening with dewdrops, nestled besides a rustic, wooden crate full of freshly harvested vegetables.

Table 1. We show examples for the three synthesis templates. Such examples are used as demonstrations for Llama-2 to perform the in-context learning task. We have 176 such examples in total. Most of them are generated by prompting GPT-4 [58], while a handful of others are human generated (in a 10M scale pilot study of synthetic captions, we do not notice significant differences between including or excluding human generated examples.)

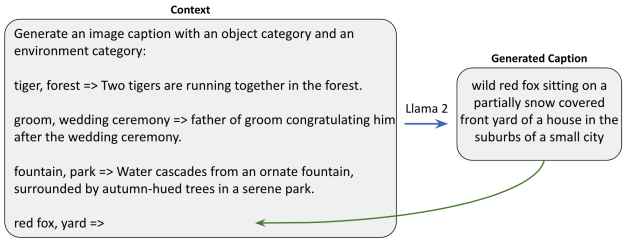

Figure 3. In-context caption generation using Llama-2 [83]. We randomly sample three in-context examples for each inference run.

of, our objective is to prompt the LLM to create captions such as *a cute yellow cat is enjoying the fish in front of the sofa*. To add variety, we have a selection of 10 different positional relationship words that we randomly choose from.

For each of the three templates, we have prepared multiple demonstration examples that serve as instructions for the LLM to complete the caption synthesis task. Table 1 shows a couple of examples for each template. In total, we have 106 examples for c–>*prompt*, 50 examples for c, bg–>*prompt*, and 20 examples for c, rel–>*prompt*. Such examples are mostly collected by prompting GPT-4, with a handful from human. In a pilot study, we do not observe difference between including or excluding human generated examples.

In the stage of generating captions in-context, we select a concept and one of the three templates. Next, we randomly pick three examples from the chosen template and frame the caption generation as a text completion task. This process is illustrated in Figure 3.

3.2. Synthesizing Images

For each text caption, we generate a variety of images by initiating the reverse diffusion process with different random noise. The Classifier-Free Guidance (CFG) scale is a crucial factor in this process. A higher CFG scale enhances the quality of the samples and the alignment between text and image, whereas a lower scale results in more diverse samples and better adherence to the original conditional distribution of images based on the given text. Following the approach used in StableRep [81], we opt for a lower CFG scale, specifically 2.5, and produce 4 images for each caption. Examples of these images can be seen in Figure 4.

3.3. Representation Learning

Our representation learning method is built upon StableRep [81]. The key component of our approach is the multi-positive contrastive learning loss [43] which works by aligning (in the embedding space) images generated from the same caption. We additionally combine multiple techniques from other self-supervised learning methods, including a patch-level masked image modeling objective. We briefly review StableRep and elaborate on the added modules.

StableRep [81] minimizes the cross-entropy loss between a ground-truth assignment distribution and a contrastive assignment distribution. Consider an encoded anchor sample $\boldsymbol{a}$ and a set of encoded candidates $\{\boldsymbol{b}_1, \boldsymbol{b}_2, ..., \boldsymbol{b}_K\}$. The contrastive assignment distribution $\mathbf{q}$ describes how likely the model predicts $\boldsymbol{a}$ and each $\boldsymbol{b}$ to be generated from the same caption, and the ground-truth distribution is the actual match

Figure 4. Random examples of synthetic captions and images generated in our SynCLR pipeline. Each caption comes with 4 images.

between $\boldsymbol{a}$ and $\boldsymbol{b}$ ($\boldsymbol{a}$ is allowed to match multiple $\boldsymbol{b}$):

$$\mathbf{q}_i = \frac{\exp(\boldsymbol{a} \cdot \boldsymbol{b}_i/\tau)}{\sum_{j=1}^{K} \exp(\boldsymbol{a} \cdot \boldsymbol{b}_j/\tau)} \tag{1}$$

$$\mathbf{p}_i = \frac{\mathbb{1}_{\text{match}(\boldsymbol{a},\boldsymbol{b}_i)}}{\sum_{j=1}^{K} \mathbb{1}_{\text{match}(\boldsymbol{a},\boldsymbol{b}_j)}} \tag{2}$$

where $\tau \in \mathcal{R}_+$ is the scalar temperature, $\boldsymbol{a}$ and all $\boldsymbol{b}$ have been ℓ_2 normalized, and the indicator function $\mathbb{1}_{\text{match}(\cdot,\cdot)}$ indicates whether two samples are from the same caption. The contrastive loss for $\boldsymbol{a}$ is given as

$$\mathcal{L}(\boldsymbol{a}) = H(\mathbf{p}, \mathbf{q}) = -\sum_{i=1}^{K} \mathbf{p}_i \log \mathbf{q}_i \tag{3}$$

iBOT [98] is a masked image modeling objective, wherein a localized patch is masked, and the model is tasked with predicting the tokenized representation of said masked patch. It adapts the DINO [11] objective from the image level into the patch level. We follow [67] to replace the softmax-centering method with the iterative Sinkhorn-Knopp (SK) algorithm [19]. We run SK for 3 iterations to build the prediction target.

Exponential Moving Average (EMA) is firstly introduced into self-supervised learning by MoCo [38]. We use EMA to encode crops as $\boldsymbol{b}$ and to produce the targets for iBOT loss. We update the EMA model as $\theta_{ema} \leftarrow \lambda\theta_{ema} + (1-\lambda)\theta$, following a cosine schedule for λ from 0.994 to 1 during training [34, 59]. We find the EMA module not only increases the final performance, but also improves the training stability for long training schedules.

Multi-crop strategy is introduced by [10] as a smart way to improve computation efficiency, and is adopted in this paper. For these local crops, we only employ the contrastive loss, omitting the iBOT loss. Local crops are encoded only by the student network, and matched to global crops from the same caption encoded by the EMA model. Such reuse of global crops saves computation. For each image x, where we generate a single global crop x^g alongside n local crops x^l, the final loss can be expressed as follows:

$$\mathcal{L}(x^g) + \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}(x_i^l) + \mathcal{L}^{iBOT}(x^g) \tag{4}$$

3.4. Implementation

Concept list. We concatenate class names from various datasets, including IN-1k [21], IN-21k (we keep the most frequent 13k classes), Aircraft [51], Cars [44], DTD [17], Flowers [55], Pets [60], Sun397 [88], Caltech-101 [30], Food-101 [7], and Places-365 [96]. If the concept is a place (*i.e.* SUN397 and Places) or a texture (*i.e.* DTD), we only apply the c –> *caption* template. For fine-grained classes such as pets or flowers, we employ GPT-4 to generate a consolidated list of probable backgrounds, rather than producing distinct lists for each specific class. We favor more frequent sampling from IN-1k, Food101, Cars, Aircraft, and Flowers.

Batches. For each training batch, we sample 2048 captions (except when noted), and use all of the 4 images generated by each caption. We generate 1 global and 4 local crops for each image. As a result, each batch contains 8192 global crops, which is similar with prior work [13, 14, 34, 81].

Masking. For the iBOT loss, we randomly choose 50% images inside a batch to mask, and randomly mask 50% of the tokens in each chosen image. We use 65536 prototypes. While the target from the EMA model is ascertained using the SK algorithm, we apply softmax normalization to the output of the student model.

Projection heads. We follow the design in MoCo v3 [14] and DINO [11] for the contrastive and iBOT loss heads, respectively, ensuring consistency with established methods.

Other hyper-parameters. We set the temperature in the contrastive loss to 0.08. For the temperature used in the iBOT loss, we linearly increase it from 0.04 to 0.07 over 4000 iterations, and keep it as 0.07 afterwards, as in DINO [11]. Additionally, the weight decay parameter is incrementally adjusted from 0.04 to 0.2, adhering to a cosine schedule.

4. Experiment

We first perform an ablation study to evaluate the efficacy of various designs and modules within our pipeline. Then we proceed to scale up the volume of synthetic data.

4.1. Study different components

We analyze each component of SynCLR, and ablate their effectiveness in two measurements: (1) linear probing performance on IN-1k; (2) average accuracy of linear transfer on

captions	StableRep		SynCLR	
	IN	avg.	IN	avg.
cc12m	73.0	81.6	77.1	85.3
IN+h+Places	75.4	80.0	78.7	83.0
IN+Places+LLM	73.7	76.9	77.6	81.8
IN+OurBG+LLM	75.3	78.5	78.2	81.9
our final config.	**75.8**	**85.7**	**78.8**	**88.1**

Table 2. **Comparison of different caption synthesis strategies**. We report top-1 ImageNet linear evaluation accuracy and the average accuracy over 9 fine-grained datasets. Every item here includes 10M captions and 4 images per caption.

CFG	2	3	4
IN top-1	72.8	72.6	72.6

Table 3. **Classifier-free guidance scale (CFG).** Contrastive loss prefers small CFG scale but is not very sensitive to it.

fine-grained datasets Aircraft [51], Cars [44], DTD [17], Flowers [55], Pets [60], Sun397 [88], Caltech-101 [30], Food-101 [7], and Pascal VOC [25]. For analysis conducted in this subsection, we train ViT-B/16 [24] models for 85000 iterations, and use the `cls` token as image representation.

Synthesize captions. Following [81], we use cc12m [12] real captions as our baseline, which has 10M sentences. To synthesize captions, we design the following variants: (a) *IN+h+Places* randomly combines one IN class plus its hypernyms in WordNet graph, with one place class; (b) *IN+Places+LLM* uses the c, bg –> *caption* in-context synthesis template with c from IN and bg from places; (c) *IN+ourBG+LLM* uses the background classes output by GPT-4, instead of Places; (d) *ours* means our full configuration specified in Section 3.1. For each of the config, we generate 10M captions. If not enough, we do duplication.

Results are summarized in Table 2, where we train both StableRep and SynCLR to avoid biases favored by a single method. Compared to a real caption dataset cc12m, simply concatenating IN and Places class names improves the ImageNet linear accuracy but reduces the fine-grained classification performance. Interestingly, naively asking Llama to combine IN and Places classes into captions yields the worst performance. Replacing random background from places with GPT generated background improves the accuracy. This shows the importance of synthesizing captions that follow the distribution of real captions, which were used to train the text-to-image model. Finally, our full configuration achieves the best accuracy on both ImageNet and fine-grained classification. Another advantage of our synthesis method is its scalability – scale up to hundreds of millions of captions with little duplication. In contrast, if we concatenate IN classes with Places classes, there are at most 365k unique captions.

Synthesize images. There are two major parameters in this process: number of images per caption and classifier free guidance scale. For the former, we find generating 4 images is almost able to reproduce StableRep [81]'s performance (10 images) when using cc12m captions (ours 73.0% v.s. StableRep 73.5% on ImageNet). Thus we stick to 4. For guidance scale, we briefly find the contrastive loss is not very sensitive to CFG in a pilot study, as shown in Table 3. Thus we stick to 2.5, similar as StableRep [81].

method	EMA	iBOT	MC	IN	avg.	ADE20k
StableRep				75.8	85.7	-
	✓			76.7	86.7	48.0
	✓	✓		77.6	87.1	50.5
	✓		✓	78.6	87.8	49.5
SynCLR	✓	✓	✓	78.8	88.1	50.8

Table 4. **Important components for our model.** ViT-B/16 models are trained for 85000 iterations. We study the modules that affect the ImageNet linear evaluation, the fine-grained classification (avg.), and ADE20k segmentation.

method	IN	avg.
Supervised CE	71.9	75.0
SimCLR	63.6	67.9
SynCLR	**75.3**	**78.5**

Table 5. **Comparison of different learning objectives.** These objectives assume different level of classification granularity, as shown in Figure 2. Our modeling, *i.e.*, defining classes as captions, outperforms the other two. To accomondate Supervised CE training, all items here used *IN+OurBG+LLM* entry in Table 2.

Model components. We present the improvement of accuracy brought by different modules in Table 4. Compared to the baseline StableRep, adding a teacher EMA model improves the IN linear accuracy by 0.9%. Further adding iBOT local objective or the multi-crop strategy increases the accuracy by 0.9% and 1.9%, respectively. Combining all of them results in our full SynCLR model, which achieves 78.8% top-1 IN linear accuracy. The fine-grained classification performance follows a similar trend, and reaches 88.1%. Besides, we test the transfer ability to semantic segmentation on ADE20k. The iBOT objective brings 1.0 more mIoU than multi-crop strategy, demonstrating the effectiveness of masked image modeling for dense prediction tasks.

Compare to SimCLR and supervised training. We compare the three different representation learning objectives shown in Figure 2, which classify images at different levels of granularity. Since supervised cross-entropy training requires a fixed set of balanced classes (indeed both *fixed set* and *balance* are limitations of such method), we use the *IN+ourBG+LLM* configuration where we have 1000 balanced classes (*i.e.*, each class has 40k images). The supervised training recipe follows [76]. For a fair comparison with SimCLR, we remove all unmatched modules (*i.e.*, EMA, iBOT, and MC) to make sure that the only difference between SimCLR and our SynCLR is the classification gran-

	text	img	# imgs		ImageNet	Aircraft	Cars	DTD	Flowers	Pets	SUN397	Caltech-101	Food-101	VOC2007	Average
StableRep	real	syn	100M	ViT-B/16	75.7	59.2	83.5	80.1	97.3	88.3	74.3	94.7	85.1	87.9	83.4
CLIP	real	real	400M	ViT-B/16	80.2	59.5	86.7	79.2	98.1	93.1	78.4	94.7	92.8	89.2	85.7
				ViT-L/14	83.9	69.4	90.9	82.1	99.2	95.1	81.8	96.5	95.2	89.6	88.9
OpenCLIP	real	real	400M	ViT-B/16	78.9	61.1	92.3	81.9	98.2	91.5	77.9	95.2	90.9	88.0	86.3
			400M	ViT-L/14	82.3	67.1	94.0	83.6	98.8	92.5	81.0	96.4	93.4	88.8	88.4
			2B	ViT-L/14	83.4	71.7	95.3	85.3	99.0	94.2	82.2	97.5	94.1	88.9	89.8
DINO v2*	-	real	142M	ViT-B/14	**83.9**†	79.4	88.2	83.3	99.6	96.2	77.3	96.1	92.8	88.2	**89.0**
				ViT-L/14	**85.7**†	81.5	90.1	84.0	99.7	96.6	78.7	97.5	94.3	88.3	90.1
SynCLR	syn	syn	600M	ViT-B/16	80.7	81.7	93.8	79.9	99.1	93.6	76.2	95.3	91.6	89.4	**89.0**
				ViT-L/14	83.0	85.6	94.2	82.1	99.2	94.1	78.4	96.1	93.4	90.3	**90.4**

Table 6. **Comparison on ImageNet linear evaluation and fine-grained classificaton.** SynCLR achieves comparable results with OpenAI's CLIP and DINO v2 models, despite *only* using synthetic data. *DINO v2 modes are distilled from a ViT-g model, thus advantageous in this comparison. † we rerun only using `cls` token instead of concatenating multiple layers presented in the original DINO v2 paper [59].

ularity defined by the contrastive loss. For all of them, we do pre-training and then linear probing on the target dataset.

Table 5 presents the comparison. Our multi-positive objective, which defines images as the same class if they are generated by the same caption, achieves the best performance. It outperforms supervised cross-entropy training and SimCLR by 3.4% and 11.7% for top-1 accuracy on ImageNet linear evaluation, and by 3.5% and 10.6% on fine-grained classification tasks. Besides, our objective does not require balance between samples from a fixed set of classes, making it easier to scale up.

4.2. Scaling up

After we have ablated different components, we scale up our experiments. Specifically, we synthesize a dataset of 150M captions, called *SynCaps-150M*, from which we generate 600M images. We train both ViT-B/16 and ViT-L/14 (no SwiGLU [73] or LayerScale [82]), and extend the training schedules to 500k steps with a batch size of 8192 captions. We use 224x224 resolution for all pre-training tasks.

We compare SynCLR with OpenAI's CLIP [62], OpenCLIP [16], and DINO v2 [59], which represent *learning from data*. We note that ViT-B/14 and ViT-L/14 from DINO v2 are distilled from a ViT-g [93] model, which makes DINO v2 advantageous in our comparison. We also includes StableRep [81], which uses the *hybrid* paradigm.

ImageNet linear evaluation. For fair comparison, `cls` token from the last block is used as representation across all models (whereas in DINO v2, results are from concatenating multiple layers). As shown in Table 6, SynCLR achieves 80.7% with ViT-B and 83.0% with ViT-L. This is similar as CLIP, but still lags behind DINO v2 by 3.2% and 2.7%, respectively, partially because of the extra distillation in DINO v2. We note SynCLR has already outperformed other self-supervised methods pre-trained directly on ImageNet-1k (*e.g.*, DINO achieves 78.2% with ViT-B/16 and iBOT reaches 81.0% with ViT-L/16).

method	pre-train data	distill	ViT-B	ViT-L
StableRep	hybrid, 100M		49.4	-
MoCo v3	real, IN1K-1M		47.3	49.1
BEiT	real, IN1K-1M+DALLE		47.1	53.3
MAE	real, IN1K-1M		48.1	53.6
iBOT	real, IN1K-1M		50.0	-
CLIP	real, WIT-400M		52.6	-
BEiT v2	real, WIT-400M, IN1K	✓	53.1	56.7
DINO v2	real, LVD-142M	✓	**54.4** †	57.5†
SynCLR	synthetic, 600M		**54.3**	**57.7** †

Table 7. **ADE20K semantic segmentation** (mIoU) using UperNet, with single scale at 512x512 resolution. † use patch size of 14x14, thus adapt to 518x518 resolution.

Fine-grained classification. On the nine fine-grained datasets we have evaluated in Table 6, SynCLR achieves very similar average accuracy as DINO v2, *e.g.*, 89.0% v.s. 89.0% for ViT-B, and 90.1% vs 90.4% for ViT-L. Both SynCLR and DINO v2 have curated the pre-training data to include the distribution for these datasets (but in different ways and portions), and end up with similar performance. Interestingly, SynCLR outperforms others on Aircraft and Cars, possibly because we favor more frequent sampling towards them. This can be an advantage for synthetic data when we know what downstream tasks to solve. Besides, SynCLR outperforms CLIP and StableRep by 3.3% and by 5.6% for ViT-B, respectively.

Semantic segmentation. To evaluate the pixel-level understanding ability of SynCLR, we fine-tune the pre-trained models on ADE20k [97], following the setup in [5, 39]. UperNet [89] is used as the task layer, and we evaluate with a single-scale, *i.e.* 512x512. Besides CLIP and DINO v2, we also compare to self-supervised methods pre-trained on ImageNet, as well as BEiT v2 [61], which distills from CLIP. Table 7 shows that our SynCLR outperforms self-supervised

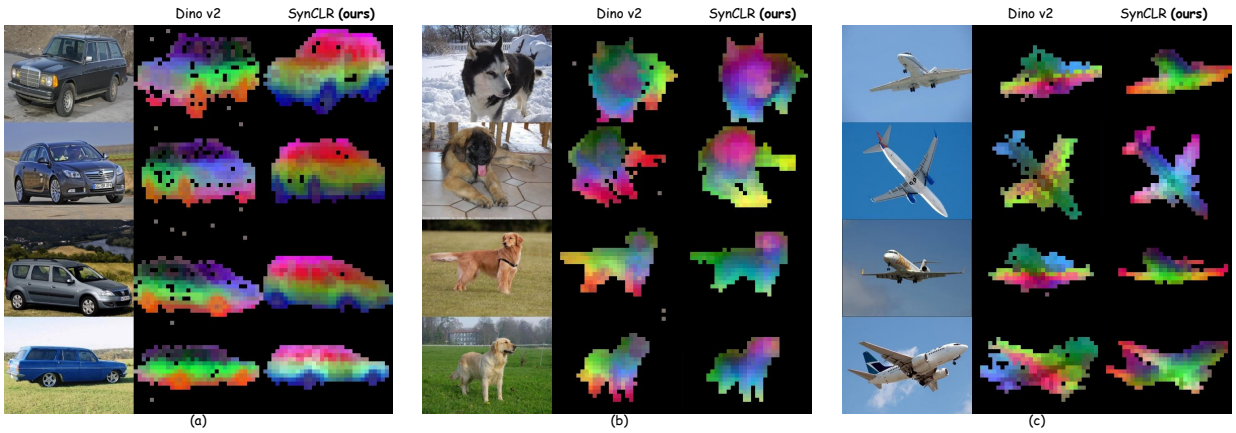

Figure 5. **PCA visualization.** Follow DINO v2 [59], we compute a PCA between the image patches from the same set and colorize by their first 3 components. Compared to DINO v2, SynCLR produces more accurate maps for cars (*e.g.*, zoom-in to see the two bars on the roof of the first car, and the three side windows of the third car) and airplanes (*e.g.*, the boundaries), while being slightly worse for dogs (*e.g.*, heads). We use ViT-L/14 for both methods. Images are resized to 336x448 resolution, yielding 24x32 visualization grids.

methods trained on IN-1k by a clear marge, *e.g.*, 4.3 higher mIoU than iBOT. Despite not involving a high resolution pre-training period like DINO v2 (*e.g.*, 518x518), SynCLR performs similarly with DINO v2 (0.1 lower for ViT-B possibly because DINO v2 uses a smaller patch size of 14x14, but 0.2 higher for ViT-L). This suggests SynCLR pre-training is suitable for dense prediction tasks.

ImageNet fine-tuning. We evaluate the fine-tuning transfer ability of SynCLR on ImageNet. Our SynCLR achieves 87.9% top-1 accuracy with ViT-L, outperforming models trained on ImageNet images or large scale image datasets. Specifically, SynCLR outperforms OpenCLIP ViT-L (87.1% top-1) trained on Laion-2B, which is the dataset Stable Diffusion (the text2image model we used) is trained on. This contrasts with [26, 69], which shows that directly training a classifier on synthetic images yields bad classification accuracy. Our finding suggests synthetic images are good for training representations, which later can be easily adapted to a downstream task with limited amount of real data. Detailed comparisons are provided in Appendix C.

PCA visualization. Following the method used in DINO v2 [59], we present visualizations derived from the Principal Component Analysis (PCA) conducted on patch features extracted using our model SynCLR. As depicted in Figure 5, a comparative analysis is conducted between SynCLR and DINO v2, both utilizing the ViT-L/14 architecture. The results demonstrate that SynCLR effectively accentuates the features of cars and planes, while efficiently minimizing background clutter.

5. Discussions and Conclusion

Why learn from generative models? One compelling reason is that *a generative model can act like hundreds of datasets simultaneously*. Traditionally, researchers have to spend separate effort collecting datasets for different image categories, *e.g.*, cars, flowers, cats, dogs, and so on. DINO v2 [59] achieves robust representations by curating and amalgamating numerous such datasets. Such a process introduces complexities such as clustering and search challenges. In contrast, advanced text-to-image generative models like Stable Diffusion [63] or Imagen [68] have the capability to generate *many* diverse datasets. These models provide the flexibility to produce an infinite number of samples (albeit finite diversity) and control the generation process through textual input. Thus, generative models offer a convenient and effective method for *curating* training data. In our study, we harness this advantage to synthesize images encompassing a broad spectrum of visual concepts.

What can be further improved? Enhanced caption sets can be achieved through various methods, such as enriching the set of in-context examples, optimizing the sampling ratios among different concepts, and utilizing more advanced LLMs. In terms of the learning process, one approach is to distill knowledge from a larger model, and incorporate an additional high-resolution training phase (as discussed in [59]) or an intermediate IN-21k fine-tuning stage (as per [5, 61]). Regarding architectural improvements, the integration of SwiGLU and LayerScale, coupled with superior model initialization strategies (referenced in [28]), can be beneficial. However, due to limited resources and the scope of this paper not being focused on achieving the highest possible metrics, we propose these areas for further exploration in future research endeavors.

In summary, this paper studies a new paradigm for visual representation learning – *learning from generative models*. Without using any real data, SynCLR learns visual representations that are comparable with those achieved by state of the art general-purpose visual representation learners.

References

[1] Hassan Abu Alhaija, Siva Karthik Mustikovela, Lars Mescheder, Andreas Geiger, and Carsten Rother. Augmented reality meets computer vision: Efficient data generation for urban driving scenes. *IJCV*, 2018. 3

[2] Mahmoud Assran, Mathilde Caron, Ishan Misra, Piotr Bojanowski, Florian Bordes, Pascal Vincent, Armand Joulin, Mike Rabbat, and Nicolas Ballas. Masked siamese networks for label-efficient learning. In *ECCV*, 2022. 3

[3] Shekoofeh Azizi, Simon Kornblith, Chitwan Saharia, Mohammad Norouzi, and David J Fleet. Synthetic data from diffusion models improves imagenet classification. *arXiv preprint arXiv:2304.08466*, 2023. 2

[4] Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. Data2vec: A general framework for self-supervised learning in speech, vision and language. In *ICML*, 2022. 3

[5] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021. 3, 7, 8

[6] Suzanna Becker and Geoffrey E Hinton. Self-organizing neural network that discovers surfaces in random-dot stereograms. *Nature*, 1992. 3

[7] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *ECCV*, 2014. 5, 6

[8] Emmanuel Asiedu Brempong, Simon Kornblith, Ting Chen, Niki Parmar, Matthias Minderer, and Mohammad Norouzi. Denoising pretraining for semantic segmentation. In *CVPR*, 2022. 3

[9] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *NeurIPS*, 2020. 3

[10] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *NeurIPS*, 2020. 5

[11] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021. 3, 5

[12] 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 *CVPR*, 2021. 6

[13] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020. 2, 3, 5

[14] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *ICCV*, 2021. 5

[15] Yuhua Chen, Wen Li, Xiaoran Chen, and Luc Van Gool. Learning semantic segmentation from synthetic data: A geometrically guided input-output adaptation approach. In *CVPR*, 2019. 3

[16] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *CVPR*, 2023. 7

[17] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *CVPR*, 2014. 5, 6

[18] Kevin Clark and Priyank Jaini. Text-to-image diffusion models are zero-shot classifiers. *arXiv preprint arXiv:2303.15233*, 2023. 3

[19] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *NeurIPS*, 2013. 5

[20] Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, Andreas Peter Steiner, Mathilde Caron, Robert Geirhos, Ibrahim Alabdulmohsin, et al. Scaling vision transformers to 22 billion parameters. In *ICML*, 2023. 3

[21] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 1, 3, 5

[22] Jeff Donahue and Karen Simonyan. Large scale adversarial representation learning. *NeurIPS*, 2019. 3

[23] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *ICML*, 2014. 3

[24] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2, 6

[25] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 2010. 6

[26] Lijie Fan, Kaifeng Chen, Dilip Krishnan, Dina Katabi, Phillip Isola, and Yonglong Tian. Scaling laws of synthetic images for model training ... for now. *arXiv:2312.04567*, 2023. 2, 3, 8

[27] Lijie Fan, Dilip Krishnan, Phillip Isola, Dina Katabi, and Yonglong Tian. Improving clip training with language rewrites. In *NeurIPS*, 2023. 3

[28] Yuxin Fang, Quan Sun, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva-02: A visual representation for neon genesis. *arXiv preprint arXiv:2303.11331*, 2023. 8

[29] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. In *CVPR*, 2023. 3

[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. In *CVPR*, 2004. 5, 6

[31] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *ICLR*, 2018. 2

[32] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014. 3

[33] Priya Goyal, Dhruv Mahajan, Abhinav Gupta, and Ishan Misra. Scaling and benchmarking self-supervised visual representation learning. In *ICCV*, 2019. 1

[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. In *NeurIPS*, 2020. 3, 5

[35] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *CVPR*, 2006. 3

[36] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 3

[37] Kaiming He, Ross Girshick, and Piotr Dollár. Rethinking imagenet pre-training. In *ICCV*, 2019. 3

[38] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020. 3, 5

[39] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022. 3, 7

[40] Ruifei He, Shuyang Sun, Xin Yu, Chuhui Xue, Wenqing Zhang, Philip Torr, Song Bai, and Xiaojuan Qi. Is synthetic data from generative models ready for image recognition? *arXiv preprint arXiv:2210.07574*, 2022. 2, 3

[41] Ali Jahanian, Xavier Puig, Yonglong Tian, and Phillip Isola. Generative models as a data source for multiview representation learning. *arXiv preprint arXiv:2106.05258*, 2021. 2, 3

[42] 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 *ICML*, 2021. 3

[43] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *NeurIPS*, 2020. 2, 4

[44] Jonathan Krause, Jia Deng, Michael Stark, and Li Fei-Fei. Collecting a large-scale dataset of fine-grained cars. *tech report*, 2013. 5, 6

[45] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NeurIPS*, 2012. 3

[46] Varun Kumar, Ashutosh Choudhary, and Eunah Cho. Data augmentation using pre-trained transformer models. *arXiv preprint arXiv:2003.02245*, 2020. 3

[47] Alexander C Li, Mihir Prabhudesai, Shivam Duggal, Ellis Brown, and Deepak Pathak. Your diffusion model is secretly a zero-shot classifier. *arXiv preprint arXiv:2303.16203*, 2023. 3

[48] Tianhong Li, Huiwen Chang, Shlok Mishra, Han Zhang, Dina Katabi, and Dilip Krishnan. Mage: Masked generative encoder to unify representation learning and image synthesis. In *CVPR*, 2023. 3

[49] Yanghao Li, Saining Xie, Xinlei Chen, Piotr Dollar, Kaiming He, and Ross Girshick. Benchmarking detection transfer learning with vision transformers. *arXiv preprint arXiv:2111.11429*, 2021. 3

[50] Hao Liu, Tom Zahavy, Volodymyr Mnih, and Satinder Singh. Palm up: Playing in the latent manifold for unsupervised pretraining. *arXiv preprint arXiv:2210.10913*, 2022. 3

[51] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv:1306.5151*, 2013. 5, 6

[52] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *CVPR*, 2016. 3

[53] Yu Meng, Jiaxin Huang, Yu Zhang, and Jiawei Han. Generating training data with language models: Towards zero-shot language understanding. *arXiv preprint arXiv:2202.04538*, 2022. 3

[54] Masato Mimura, Sei Ueno, Hirofumi Inaguma, Shinsuke Sakai, and Tatsuya Kawahara. Leveraging sequence-to-sequence speech synthesis for enhancing acoustic-to-word speech recognition. In *SLT*, 2018. 3

[55] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics & Image Processing*, 2008. 5, 6

[56] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *ECCV*, 2016. 2

[57] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 3

[58] OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 3, 4

[59] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 1, 2, 3, 5, 7, 8

[60] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *CVPR*, 2012. 5, 6

[61] Zhiliang Peng, Li Dong, Hangbo Bao, Qixiang Ye, and Furu Wei. Beit v2: Masked image modeling with vector-quantized visual tokenizers. *arXiv preprint arXiv:2208.06366*, 2022. 7, 8

[62] 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 *ICML*, 2021. 1, 2, 3, 7

[63] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 8

[64] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 3

[65] Andrew Rosenberg, Yu Zhang, Bhuvana Ramabhadran, Ye Jia, Pedro Moreno, Yonghui Wu, and Zelin Wu. Speech recognition with augmented synthesized speech. In *ASRU*, 2019. 3

[66] Nick Rossenbach, Albert Zeyer, Ralf Schlüter, and Hermann Ney. Generating synthetic audio data for attention-based speech recognition systems. In *ICASSP*, 2020. 3

[67] Yangjun Ruan, Saurabh Singh, Warren Morningstar, Alexander A Alemi, Sergey Ioffe, Ian Fischer, and Joshua V Dillon. Weighted ensemble self-supervised learning. *arXiv preprint arXiv:2211.09981*, 2022. 5

[68] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*, 2022. 8

[69] Mert Bulent Sariyildiz, Karteek Alahari, Diane Larlus, and Yannis Kalantidis. Fake it till you make it: Learning transferable representations from synthetic imagenet clones. In *CVPR*, 2023. 2, 3, 8

[70] Saurabh Saxena, Charles Herrmann, Junhwa Hur, Abhishek Kar, Mohammad Norouzi, Deqing Sun, and David J Fleet. The surprising effectiveness of diffusion models for optical flow and monocular depth estimation. *arXiv preprint arXiv:2306.01923*, 2023. 3

[71] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. In *NeurIPS*, 2022. 1

[72] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. In *CVPR workshops*, 2014. 3

[73] Noam Shazeer. Glu variants improve transformer. *arXiv preprint arXiv:2002.05202*, 2020. 7

[74] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *Nature*, 2017. 3

[75] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 3

[76] Andreas Steiner, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. How to train your vit? data, augmentation, and regularization in vision transformers. *arXiv preprint arXiv:2106.10270*, 2021. 6

[77] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. Alpaca: A strong, replicable instruction-following model. *Stanford Center for Research on Foundation Models.*, 2023. 3

[78] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. *arXiv:1906.05849*, 2019. 3

[79] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? In *NeurIPS*, 2020. 3

[80] Yonglong Tian, Olivier J Henaff, and Aäron van den Oord. Divide and contrast: Self-supervised learning from uncurated data. In *ICCV*, 2021. 1

[81] Yonglong Tian, Lijie Fan, Phillip Isola, Huiwen Chang, and Dilip Krishnan. Stablerep: Synthetic images from text-to-image models make strong visual representation learners. In *NeurIPS*, 2023. 1, 2, 3, 4, 5, 6, 7

[82] Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. Going deeper with image transformers. In *ICCV*, 2021. 7

[83] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 3, 4

[84] Gul Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *CVPR*, 2017. 3

[85] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *ICML*, 2020. 3

[86] Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. In *CVPR*, 2022. 3

[87] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *CVPR*, 2018. 3

[88] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, 2010. 5, 6

[89] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *ECCV*, 2018. 7

[90] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *CVPR*, 2022. 3

[91] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *CVPR*, 2023. 3

[92] Yiben Yang, Chaitanya Malaviya, Jared Fernandez, Swabha Swayamdipta, Ronan Le Bras, Ji-Ping Wang, Chandra Bhagavatula, Yejin Choi, and Doug Downey. Generative data augmentation for commonsense reasoning. *arXiv preprint arXiv:2004.11546*, 2020. 3

[93] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *CVPR*, 2022. 7

[94] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *ECCV*, 2016. 2

[95] Wenliang Zhao, Yongming Rao, Zuyan Liu, Benlin Liu, Jie Zhou, and Jiwen Lu. Unleashing text-to-image diffusion models for visual perception. *arXiv preprint arXiv:2303.02153*, 2023. 3

[96] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *CVPR*, 2016. 3, 5

[97] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *IJCV*, 2019. 7

[98] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. *arXiv preprint arXiv:2111.07832*, 2021. 2, 3, 5

[99] Yongchao Zhou, Hshmat Sahak, and Jimmy Ba. Training on thin air: Improve image classification with generated data. *arXiv preprint arXiv:2305.15316*, 2023. 3