Robust Cross-Modal Representation Learning with Progressive Self-Distillation

Alex Andonian*
MIT CSAIL
andonian@mit.edu

Shixing Chen, Raffay Hamid
Amazon Prime Video
{shixic, raffay}@amazon.com

Abstract

The learning objective of vision-language approach of CLIP [63] does not effectively account for the noisy many-to-many correspondences found in web-harvested image captioning datasets, which contributes to its compute and data inefficiency. To address this challenge, we introduce a novel training framework based on cross-modal contrastive learning that uses progressive self-distillation and soft image-text alignments to more efficiently learn robust representations from noisy data. Our model distills its own knowledge to dynamically generate soft-alignment targets for a subset of images and captions in every minibatch, which are then used to update its parameters. Extensive evaluation across 14 benchmark datasets shows that our method consistently outperforms its CLIP counterpart in multiple settings, including: (a) zero-shot classification, (b) linear probe transfer, and (c) image-text retrieval, without incurring extra computational cost. Analysis using an ImageNet-based robustness test-bed [70] reveals that our method offers better effective robustness to natural distribution shifts compared to both ImageNet-trained models and CLIP itself. Lastly, pretraining with datasets spanning two orders of magnitude in size shows that our improvements over CLIP tend to scale with number of training examples.

1. Introduction

The convergence of self-supervised pretraining techniques in natural language processing and computer vision have brought about a renaissance of cross-modal representation learning methods [1, 19, 30, 39, 52, 63, 68, 75] where large-scale weakly correlated multimodal data (*e.g.,* image-text pairs) is used to learn cross-modal representations using contrastive learning techniques. In particular, the recently proposed CLIP [63] model has garnered significant attention due to its impressive zero-shot recognition ability and excellent transfer performance on downstream tasks.

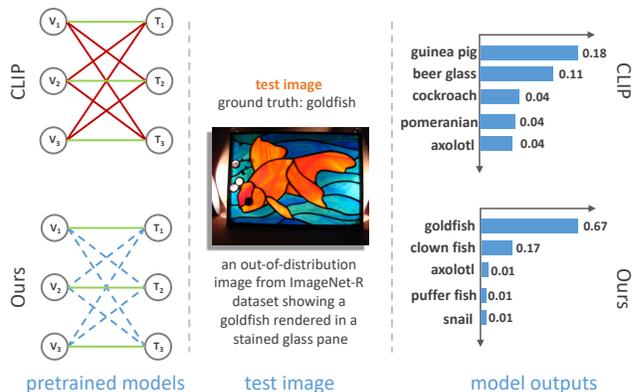However, despite their recent success, multimodal pretraining methods like CLIP [63] are data and compute in-

---

Figure 1. **Illustrative Comparison with CLIP –** CLIP [63] learns a joint vision-language embedding space by bringing corresponding image-text representation together (green links), while repelling unpaired instances away from each other (red links). This formulation does not account for potential semantic similarity between negative samples. We address this issue by learning to predict a distribution of soft-alignment targets (dotted blue edges) in a given minibatch, thereby allowing our model to learn more robust representations. This robustness is evident when comparing predicted distributions on an out-of-distribution image from ImageNet-R dataset [25], where unlike CLIP, our method can correctly classify a goldfish rendered in a stained glass pane.

efficient. Much of CLIP's success can be attributed to its voracious appetite for training data, utilizing 400M image-text pairs and an estimated 3, 584 GPU days for pretraining. As the scale of data increases, pretraining requirements of these methods become increasingly expensive, thereby limiting their widespread adoption in a sustainable manner.

This data and compute inefficiency of CLIP [63] can be partially attributed to the underlying assumptions it makes about the web-harvested data it uses for training. Several mainstream vision-language datasets utilize the alt-text HTML attribute of images scraped from archived web pages [9, 67, 71] where captions can often have words unrelated to their corresponding image-content [67]. However, CLIP [63] models the caption for each image to be accurately and exclusively related to only that image (see Figure 1). Moreover, when using larger batch sizes (32K used for CLIP), the likelihood of observing negatives with

high semantic similarity increases which can further degrade the learned representations especially those associated with shared semantics between faulty negatives [2].

To address this challenge, we propose to model the many-to-many relationships between images of web-harvested datasets and their corresponding captions more accurately using *soft probabilities* rather than hard pairing labels. Specifically, we propose a simple yet effective framework for robust contrastive language-image pretraining that uses progressive self-distillation and soft image-text alignment targets to more efficiently learn from noisy data. Instead of explicitly finding, correcting or even pruning noisy correspondences [75, 88], our joint student-teacher model dynamically generates a new set of soft-alignments for a random subset of images and captions in every minibatch. This enables our method to model many-to-many relationships while simultaneously re-calibrating potentially poorly matched instances without needing to identify them. Over the course of training, our network generates soft-alignments for increasingly large subsets of a minibatch, effectively becoming its own teacher. We identify several key elements that allow the student network to predict its targets without representation collapse or reinforcing its mistakes.

We use multiple pretraining datasets to extensively compare our approach to CLIP [63] evaluated on 14 benchmark datasets, where our approach consistently outperforms CLIP under multiple settings. Analysis using an ImageNet-based robustness test-bed [70] shows that our method offers better effective robustness to natural distribution shifts compared to both ImageNet-trained models as well as CLIP. Pretraining with datasets spanning two orders of magnitude in size shows that our improvements over CLIP tend to scale with number of training examples. Lastly, the simplicity of our approach allows it to be readily incorporated into existing and future methods.

2. Related Work

a.   Self-Supervised Representation Learning: Self-supervised learning (SSL) approaches use a *pretext task* to automatically generate a supervision signal from the data itself, thereby eliminating the dependence on expensive manual data-labeling [31]. Pretext tasks in computer vision include spatial reasoning [15, 20, 32, 54, 59], temporal context [23, 32, 36, 48, 49], and other visual properties such as hue [13, 35, 81, 82], brightness [29, 72] or optical flow [29, 62, 73, 79], reconstruction of modified inputs [59, 74, 81], and classifying inputs with pseudo-labels [15, 16, 58] or pseudo-clusters [7, 8, 86, 87]. A promising subset of SSL methods uses a variant of instance discrimination framework [17, 77] which learns to align augmented versions of features while distinguishing them from features of other instances using a contrastive loss [8, 11, 24, 47].

b.   Vision-Language Pretraining: Joint vision-language pretraining (VLP) is an active research area [1, 19, 39, 63, 68] where the availability of large-scale image-text datasets *e.g.,* YFCC100M [71] and Conceptual Captions [9, 67] has played a key role in its progress. Although multiple concurrent works are being proposed to further improve VLP models [75], our work is different from them in a few important ways. Specifically, unlike EfficientCLIP [75] that proposes an ensemble approach to obtain a less noisy data subset for cross-modal training, our method attempts to sidestep this problem altogether by re-purposing as opposed to completely removing noisy data. Similarly, DeCLIP [39] improves on the data-efficiency of CLIP [63] by leveraging intra-model contrastive learning along with a nearest-neighbor feature bank to augment negatives. However, incorporating these supervision sources can be computationally expensive. In contrast, our approach offers a simple yet effective way to improve the data-efficiency of CLIP [63] without incurring additional computational cost.

c.   Learning from Noisy Data: Several techniques have been developed to increase the robustness of label noise especially in the supervised context [22, 40, 60, 65, 83]. These techniques include loss functions that reduce impact of outliers [76, 83], metalearning procedures that learn how to correct sources of label noise [1, 37, 69, 84], loss correction approaches that model label noise [50, 51, 65], regularization techniques aimed at lowering the impact of noise [61], and noise filtering processes that iteratively refine dataset labels and retrain models to obtain a more robust final model [55]. However, these works investigate noise-robust methods in the context of common object detection and classification tasks and cannot directly be applied effectively to cross-modal pretraining tasks. As it stands, noise-robust VLP pretraining methods are still a relatively unexplored topic.

d.   Knowledge Distillation: Approaches for knowledge distillation (KD) [27] aim to transfer knowledge from one model (*i.e.*, a teacher) to another (*i.e.*, a student). While KD techniques are often motivated by certain performance and efficiency goals [6, 10, 38, 66], researchers have also found that KD methods serve as an effective regularization technique that can reduce model overfitting and improve generalization capabilities [14, 40, 43, 53]. Our approach is motivated by the recent success of self-knowledge distillation approaches [21, 78, 80] that use the student network as a teacher under supervised settings to achieve high accuracy with reduced computational cost. To the best of our knowledge, we are among the first to investigate progressive self-distillation in the context of vision-language pretraining.

3. Methods
Unlike CLIP [63], we view the problem of learning aligned vision-language representations from web-scale weakly-annotated data as the challenge of learning many-to-many

relationships from noisy image-text correspondences. To address this challenge, we propose a novel vision-language pretraining method that progressively distills a model's own knowledge to soften its initially hard-target alignments, thereby enabling it to learn more transferable representations from the same amount of training data (see Figure 2). In the following, we first establish cross-modal contrastive learning objective and identify some of its limitations. We then introduce our novel progressive self-distillation approach and explain how it addresses these limitations.

3.1. Preliminaries

We consider a batch of N semantically paired image-text tuples $\{(v_i, t_i)\}_{i=1:N}$ drawn from a cross-modal dataset. The goal of cross-modal contrastive pretraining is to learn encoders f_v for image data and f_t for text data such that for a given semantically related instance (v_i, t_i), the encoded ℓ_2-normalized embeddings $\mathbf{v}_i = f_v(v_i)$ and $\mathbf{t}_i = f_t(t_i)$ with $\mathbf{v}_i, \mathbf{t}_i \in \mathbb{R}^d$ are close together (*i.e.*, "aligned") under some distance metric, while the unpaired image and text embeddings are farther apart (*i.e.*, "unaligned").

3.2. Contrastive Learning with InfoNCE Loss

Recall that CLIP [63] trains these image and text encoders with a contrastive loss by minimizing the InfoNCE [56] loss $\mathcal{L}_{\text{InfoNCE}} = \mathcal{L}_v + \mathcal{L}_t$, where $\mathcal{L}_v$ is the loss for aligning images to text and $\mathcal{L}_t$ is the loss for aligning text to images. Specifically, $\mathcal{L}_v$ is defined as:

$$\mathcal{L}_v = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{N} \mathrm{I}_{ij} \log \mathrm{P}_v(\mathbf{v}_i, \mathbf{t}_j; \tau) \quad (1)$$

$$\mathrm{P}_v(\mathbf{v}_i, \mathbf{t}_j; \tau) = \frac{\exp(\mathrm{sim}(\mathbf{v}_i, \mathbf{t}_j)/\tau)}{\sum_{k=1}^{N} \exp(\mathrm{sim}(\mathbf{v}_i, \mathbf{t}_k)/\tau)} \quad (2)$$

where $\mathrm{sim}(\mathbf{v}_i, \mathbf{t}_j) = \mathbf{v}_i^T \mathbf{t}_j$ is typically chosen to be the dot product (cosine similarity), τ is a learnable softmax temperature parameter and I_{ij} is an element from the identity matrix $\mathbf{I}_N$. Since InfoNCE is symmetric, $\mathcal{L}_t$ and P_t are defined in a similar manner.

For convenience, let $\mathbf{V}, \mathbf{T} \in \mathbb{R}^{N \times d}$ be matrices that contain a batch of image and text embeddings whose rows are populated with $\mathbf{v}_{1:N}$ and $\mathbf{t}_{1:N}$, respectively. Then, the InfoNCE loss can be re-written compactly in matrix form as:

$$\mathcal{L}_{\text{InfoNCE}} = \mathcal{H}(\mathbf{I}_N, \rho(\mathbf{V}\mathbf{T}^\top)) + \mathcal{H}(\mathbf{I}_N, \rho(\mathbf{T}\mathbf{V}^\top)), \quad (3)$$

where $\mathcal{H}$ is the batched (row-wise) cross-entropy function with mean reduction and ρ is the standard softmax function applied row-wise such that each row sums to one.

Equation 3 shows that InfoNCE loss is simply the cross entropy between a one-hot distribution I_{ij} and estimated probability $\mathrm{P}_v(\mathbf{v}_i, \mathbf{t}_j; \tau)$. It enforces the strict assumption that an image v_i selected from a batch should be paired exclusively with text t_i while being repelled from all other t_j.

However, this assumption generally does not hold for two important reasons. First, it is likely that a given image would be aligned to several text captions to different degrees especially under large batch-size settings. Second, the ground truth pairings in large-scale weakly-annotated datasets may be simply incorrect or describe loose correlations between images and their corresponding texts.

3.3. Distillation through Soft-Alignments

To address the aforementioned limitations of using the InfoNCE loss to train on noisy cross-modal data, we propose to adopt a knowledge distillation framework where the predictive probabilities produced by a teacher network are used as soft target distributions to train a student network.

Our framework offers two key advantages. First, in the process of generating target image-text alignments, a well-trained teacher can combat poorly captioned images by re-pairing them with stronger semantic matches from the batch, thereby providing a cleaner learning signal for the student network. Second, by providing soft targets, the teacher can convey many-to-many relationships in a batch.

Specifically, to estimate the correspondence between image v_i and text t_i, our teacher model employs image and text encoders $\tilde{f}_v$ and $\tilde{f}_t$ to compute ℓ_2-normalized teacher embeddings $\tilde{\mathbf{v}}_i$ and $\tilde{\mathbf{t}}_j$ which are similarly batched as rows in the matrices $\tilde{\mathbf{V}}$ and $\tilde{\mathbf{T}}$ respectively. Our method uses a swapped prediction strategy to produce soft target distributions $\mathbf{A}^v$ and $\mathbf{A}^t$ to supervise student training. These target distributions are defined as:

$$\mathbf{A}^v = \rho(\tilde{\mathbf{T}}\tilde{\mathbf{V}}^\top; \tilde{\tau}) \text{ and } \mathbf{A}^t = \rho(\tilde{\mathbf{V}}\tilde{\mathbf{T}}^\top; \tilde{\tau}) \quad (4)$$

where ρ is the standard softmax function now using a secondary teacher temperature $\tilde{\tau}$ that transforms and re-scales raw logits into probabilities.

Swapped prediction improves on the well established forward bootstrapping approach [65] by using predictions from opposite modality. Unlike bootstrapping, swapped prediction computes the image alignment scores $\mathbf{A}^v$ from the text encoder posterior probabilities and vice-versa, thus aggregating information over all other instances from the opposite modality. Intuitively, the strength of alignment from image v_i to text t_j is based on the probability that text t_j should be matched with image v_i compared to all other v_j. This strategy has shown promise in related contrastive learning settings [50], which is consistent with our results.

Replacing the $\mathbf{I}_N$ targets in Equation 3 with estimated soft-alignment probabilities $\mathbf{A}^v$ and $\mathbf{A}^t$ allows the teacher to re-calibrate the attractive and repulsive forces between image and text embeddings in the representation space based on its estimated similarity between instances. For instance, a faulty negative pair (v_i, t_j), which may have high semantic similarity is assigned a similarity score of zero by
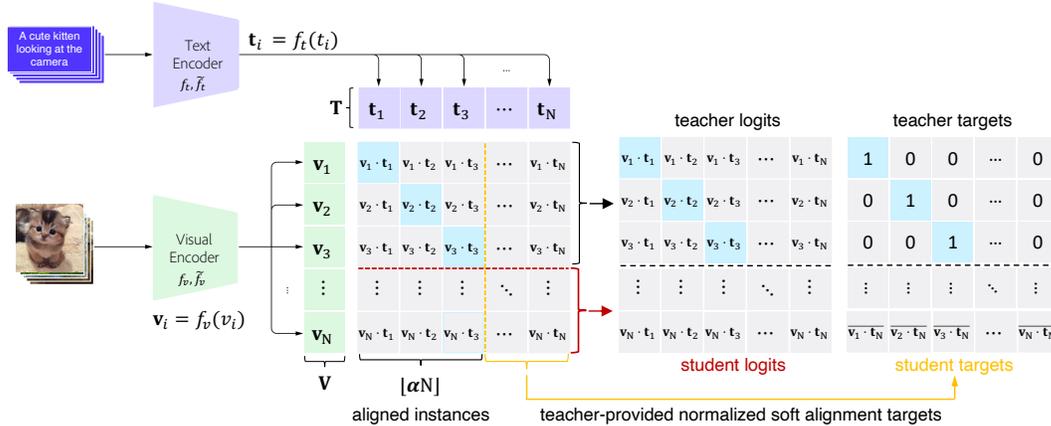
Figure 2. **Approach Overview** – We partition a minibatch into aligned instances (rows above the dotted red line) and unaligned instances (rows below the dotted red line). Our teacher network is trained on the aligned instances using the standard InfoNCE loss. By normalizing the predictions from the opposite modality (columns to the right of the dotted yellow line), the teacher provides the estimated soft-alignments for the unaligned data to supervise the student. Over the course of training, the proportion of soft-alignments provided to the student increases as the teacher's representations become more reliable.

the InfoNCE loss, whereas our method provides A_{ij} as a target which should be larger given a well trained teacher.

3.4. Progressive Self-Distillation

We now explain how to start such a teacher network and how its contribution to learning process evolves over time.

3.4.1 Teacher Network Selection

Conventional KD and SSL methods offer numerous potential teacher choices, *e.g.,* larger but static pretrained teacher networks [27], or networks that share the same model architecture but use weights from a previous epoch [42], or as an exponential moving average [24]. The primary drawback to these approaches is reduced computational and memory efficiency as they require a secondary inference stage using additional model weights that must be kept in memory.

To circumvent these issues, we adopt a *self-distillation* strategy where the student network acts as its own teacher (*i.e.,* $f_v = \tilde{f}_v, f_t = \tilde{f}_t$). The idea here is to update the targets of the student contrastive objective using the current state of the model. Intuitively, as the learning improves over time, its representation can be trusted to make more accurate predictions. This mitigates the negative effects of noisy pairings as incorrect pairs are increasingly likely to be inconsistent with the consensus learned from the rest of the data as training progresses. By refining inconsistent alignments, the model can develop more coherent representations, which further improves its ability to evaluate the consistency of noisy image-text pairs.

3.4.2 Progressing from Student to Teacher

As our objective relies on some basic level of alignment between corresponding image and text representations, we introduce a novel procedure that progressively increases the

contribution of self-distillation to the contrastive learning process over the course of training. Our model therefore dynamically evolves into its own teacher as training progresses, which differs from the standard knowledge distillation setting where the teacher is often static and separate.

We achieve this dynamic progression by randomly partitioning a batch of N image-text pairs into $N^a = \lfloor \alpha N \rfloor$ "aligned" instances and $N^u = N - \lfloor \alpha N \rfloor$ "unaligned" instances where $\alpha \in [0, 1]$ determines their relative proportions. The aligned subset is used to train the teacher network using the hard ground truth pairings and the standard InfoNCE loss. The teacher network then employs our aforementioned swapped prediction strategy to estimate soft-alignments on the unaligned instances to supervise the student. We refer to this random minibatch partitioning as *dynamic* as opposed to static as the global partitioning of instances into aligned and unaligned subsets is refreshed for each training epoch.

To increase the strength of the teacher's influence on learning, we decrease the value of α gradually, in the same way that the learning rate can be scheduled. While there are several strategies to decrease α as a function of the training iteration, *e.g.,* step-wise, linear, *etc.*, we use a cosine-annealing schedule [38] specified by a start and end value.

To summarize our overall learning procedure, we first compute the batched student-teacher embedding with $V = \tilde{V}$ and $T = \tilde{T}$. Next, we extract the first N^a rows to form aligned subset of teacher embeddings $\tilde{V}^a, \tilde{T}^a$, and the last N^u rows for the unaligned student embeddings V^u, T^u. Altogether, our final objective function is defined as:

$$\mathcal{L}_{\text{InfoNCE}}^{\text{PSD}} = \alpha \left[\mathcal{H}(\mathbf{I}_{N_a}, \rho(\tilde{V}^a \tilde{T}^\top)) + \mathcal{H}(\mathbf{I}_{N_a}, \rho(\tilde{T}^a \tilde{V}^\top)) \right] +$$
$$(1 - \alpha) \left[\mathcal{H}(\mathbf{A}^v, \rho(V^u T^\top)) + \mathcal{H}(\mathbf{A}^t, \rho(T^u V^\top)) \right]$$

where $\mathbf{I}_{N_a} \in \mathbb{R}^{N_a \times N}$ is the zero-padded identity matrix, while $\mathbf{A}^v, \mathbf{A}^t$ are indexed to match the unaligned student embeddings.

4. Experiments

We start by describing our experimental setup which aims to match CLIP [63] as closely as possible for fair comparison. We then demonstrate the advantages of our method over CLIP [63] for: (a) zero-shot classification, (b) finetuning (*i.e.*, linear probe), and (c) image-text retrieval.

4.1. Pretraining Datasets

We apply our pretraining method to three image-text datasets varying in scale, scope and noise:

a. MS COCO Captions [41] – A widely used standard image captioning benchmark dataset with approximately 118K images, each labeled with 5 human evaluated captions, and a testing set of 5K images.

b. Conceptual Captions 3M (CC3M) [67] – A collection of over 3M images and their raw descriptions harvested from the alt-text HTML attribute associated with the web-scraped images, therefore representing a wider variety of content styles. After downloading and preprocessing, we utilized about 2.9M image-text pairs in our experiments.

c. Conceptual Captions 12M (CC12M) [9] – By relaxing multiple image and text filters used in CC3M [67], CC12M is a less precise but $4\times$ larger set of image-text pairs that covers a wider range of visual concepts. Due to unavailable URLs, we utilize about 10M examples from this dataset.

4.2. Pretraining Details

In the following experiments, the image encoders follow ViT-B/32 vision transformer architecture proposed in CLIP [63], while the text encoder's transformer-based architecture follows modification proposed in [63]. Image and text features are projected to a shared 512-D space and L2 normalized before participating in contrastive loss.

Models are trained from scratch for 100 epochs using the Adam optimizer [33] with weight decay and a cosine annealing learning rate schedule with warmup [44]. As done in [63], the learnable temperature parameter τ is initialized to 0.07 and clamped to values less than 100. Automatic mixed-precision [45] training is used to save on memory and achieve minibatch sizes of 4096. Input images are randomly cropped and resized to 224×224 resolution during pretraining and the maximum length of the text is limited to 77 tokens via random sub-sequence sampling similar to [63]. Training is conducted on as many as 8 Nvidia A-100 GPUs with the longest experiments spanning up to several days. The partitioning factor α is decayed from 0.8 to 0.2 over the course of training using cosine annealing. De-

tails regarding hyperparameter values for different datasets and models are provided in the supplementary material.

We compare our approach to a re-implementation of CLIP based on the methods and pseudo-code described in the original paper as well as open source implementations such as OpenCLIP [28]. The performance of our re-implementation is consistent with that achieved by Open-CLIP and unreleased CLIP models trained on reduced subsets of the 400M dataset; thus, we retrain a baseline CLIP model for each pretraining dataset to serve as a proxy.

4.3. Evaluation Details

a. Evaluation Protocol – We measure zero-shot and linear probe classification performance using Top-1 accuracy. For the linear probe experiments, we train the linear classifier with the L-BFGS optimizer on extracted visual features as described in [63]. We use standard retrieval metrics: recall at rank K (R@K, higher is better) and mean rank (MnR, lower is better), to evaluate the retrieval performance of our model. R@K (Recall at K) calculates the percentage of test samples for which the correct result is found in the top-K retrieved points to the query sample.

b. Benchmark Datasets – We evaluate the zero-shot and linear probe classification performance of proposed approach on a suite of benchmark evaluation datasets, including ImageNet [12], Places365 [85], ObjectNet [3], and several recent variants of ImageNet aimed at evaluating the robustness of trained models to natural (as opposed to synthetic) distribution shifts [25, 26, 64].

c. Prompt Ensembling with Templates – Consistent with previous work, we find that our approach benefits from prompt ensembling to augment the original class label for downstream tasks. For fair comparison, we use the same set of prompt templates published in CLIP [63], which generally take on the form of "a photo of a {label}."

4.4. Zero-Shot Image Classification

Following the pretraining stage, we evaluate our method on zero-shot image classification achieved through natural language input, and compare to its CLIP counterpart trained under identical settings.

Table 1 lists the top-1 accuracy (%) of zero-shot image classification across a suite of benchmark datasets. Given a fixed amount of pretraining data, our method considerably outperforms its CLIP counterparts in terms of average top-1 accuracy over all datasets, achieving as much as a 6.19% absolute improvement in the CC3M pretraining regime. Notably, our method surpasses CLIP on ImageNet and all of the ImageNet variants. Our method also achieves substantial performance gains on ImageNet-Renditions (ImageNet-R) [25], a dataset specifically aimed at evaluating out-of-distribution robustness, with an average

Pretraining Dataset	Method	Cifar10 [34]	Cifar100 [34]	Caltech101 [18]	Places365 [85]	ObjectNet [3]	ImageNet-R [25]	ImageNet-O [26]	Imagenet-A [26]	ImageNetV2 [64]	ImageNet [12]	Average
COCO	CLIP	64.14	19.57	32.88	12.78	4.98	8.27	8.0	3.32	7.41	8.18	16.87
	Ours	**66.74**	**24.49**	**34.26**	**14.15**	**6.18**	**11.25**	**9.85**	**5.32**	**8.99**	**9.49**	**19.07**$^{+2.22}$
CC3M	CLIP	73.90	30.60	54.07	24.54	4.49	28.33	18.5	7.827	21.43	23.56	28.73
	Ours	**80.15**	**38.27**	**64.45**	**28.07**	**9.21**	**37.31**	**26.2**	**10.81**	**26.70**	**27.96**	**34.91**$^{+6.19}$
CC12M	CLIP	75.29	41.94	75.86	31.29	12.50	51.34	31.90	13.25	34.89	37.87	40.61
	Ours	**84.84**	**51.34**	**80.00**	**34.08**	**15.24**	**59.29**	**33.0**	**18.85**	**39.16**	**42.24**	**45.85**$^{+5.23}$

Table 1. **Zero-Shot Image Classification Comparison** – Zero-shot top-1 accuracy (%) of our method compared to baseline CLIP on numerous ImageNet-based benchmark datasets using different pretraining datasets varying in scale.
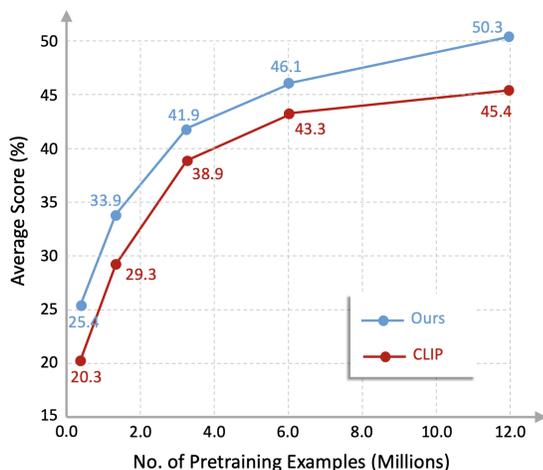

Figure 3. **Pretraining Data Size and Relative Performance** – Average zero-shot classification performance of our method compared to CLIP as number of pretraining examples from the CC12M dataset grows from 600K to 12M. Our method consistently learns representations more amenable to zero-shot classification than CLIP given same amount of pretraining data.
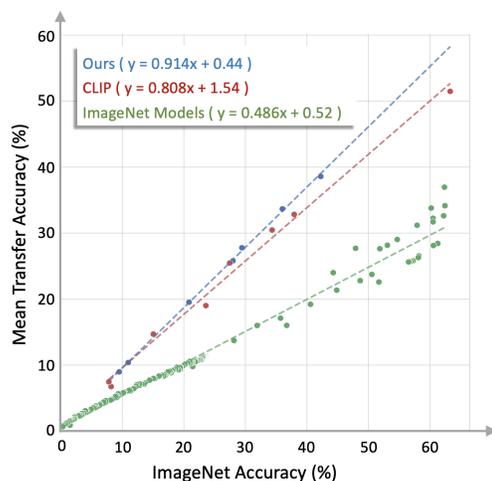

Figure 4. **Effective Robustness Evaluation** – Our method produces features more robust to natural distribution shift compared to CLIP features. When comparing models with similar ImageNet performance, our method produces representations that offer better performance on naturally shifted distributions. Mean Transfer Accuracy is computed over the ImageNet-R/O/A/V2 test sets. Best-fit trend lines suggests that the effective robustness of our method (blue) outpaces CLIP (red) as ImageNet accuracy increases.

performance gain of nearly 8.5% when considering Conceptual Captions datasets [9, 67]. Figure 1 shows how our method is able to provide more accurate predictions for a goldfish *rendered* in a stained glass pane.

To better understand the relationship between the amount of pretraining data and relative differences in downstream performance between our method and CLIP, we perform a sweep of experiments that restrict pretraining to 5, 10, 25, 50, and 100% of the CC12M dataset. Figure 3 plots the average zero-shot classification score for both methods, and shows that our method maintains higher performance throughout this range. Interestingly, our method appears to provide the largest performance increases at the boundaries of the plot, suggesting that our method can improve data-efficiency in data-limited regimes, in addition to providing larger performance gains as the dataset size increases.

4.5. Evaluation of Effective Robustness

CLIP models [63, 70] have been found to be more robust to natural distribution shifts when compared to standard models trained on ImageNet. This phenomena is illustrated in Figure 4 with ImageNet accuracy on the x-axis and mean accuracy on ImageNet-A, ImageNet-O, ImageNet-R, and ImageNetV2, on the y-axis. Previous work [46, 70] has found that the in-distribution and out-of-distribution accuracies of ImageNet models follow a predictable linear trend (plotted in green), and CLIP models established a trend (plotted in red) of improved effective robustness. Note that the slope of linear-fit to our model is higher than that for the CLIP model, suggesting that our effective robustness improves over CLIP with increasing scale.

Pretraining Dataset	Method	Food101 [5]	OxfordPets [57]	Birdsnap [4]	ImageNet [12]
COCO	CLIP	53.04	76.86	37.17	52.66
	Ours	**53.60**	**80.59**	**42.85**	**56.21**
CC3M	CLIP	53.33	78.11	37.75	56.28
	Ours	**60.69**	**80.40**	**43.17**	**61.00**
CC12M	CLIP	67.41	85.17	41.06	59.42
	Ours	**71.87**	**86.32**	**47.16**	**65.33**

Table 2. **Linear Probe Performance** – Linear probe top-1 accuracy (%) of our method compared to baseline CLIP on four benchmark datasets is given. Our method learns visual features that consistently achieve improved finetuning performance, suggesting that our loss helps to improve within-modal feature alignment.

4.6. Linear Probe Performance

We report our linear probe performance on 4 downstream datasets in Table 2. Our method outperforms its CLIP counterpart in every case, suggesting that our learned visual features alone (*i.e.,* considering within-modal alignment only) are more transferable than CLIP.

4.7. Image-Text Retrieval

Given the cross-modal nature of models under investigation, image-text retrieval consists of two sub-tasks: image-to-text and text-to-image. We evaluate our method on COCO test set with 5K images, each with 5 unique captions.

Table 3 shows our zero-shot retrieval performance relative to baseline CLIP under COCO, CC3M and CC12M pretraining datasets. Relative to CLIP, our method consistently outperforms in both text-to-image and image-to-text subtasks, with both methods generally achieving slightly higher performance in text-to-image sub-task. Note that, due to data domain differences, pretraining directly on COCO produces models with higher downstream performance than pretraining on CC12M despite the $100\times$ increase in datasize. The last four rows show that, as expected, finetuning Conceptual Captions pretrained models on COCO produces retrieval performance that exceeds corresponding baselines and the performance from training on COCO alone.

4.8. Ablation Study

We now investigate the contributions of individual components of our method, specifically the swapped prediction strategy for estimating soft-alignments over the forward bootstrapping strategy, dynamic minibatch partitioning, and progressive self-distillation to downstream performance. Following pretraining on COCO, we measure models' performance on COCO zero-shot image-text retrieval and ImageNet zero-shot classification.

	Method		Performance	
Prediction Mode	Dynamic Partitioning	Progressive Distillation	IN Top-1	MS R1
Baseline	-	-	8.18	27.58
Forward	✗	✗	6.23	20.40
Swapped	✗	✗	8.46	23.51
Forward	✓	✗	8.37	25.64
Swapped	✓	✗	8.81	26.24
Forward	✓	✓	8.87	26.18
Swapped	✓	✓	9.51	28.48

Table 4. **Abalation Study** – Our results using COCO for pretraining. IN Top-1 is zero-shot ImageNet accuracy and MS R1 is mean recall at rank 1 across image-to-text and text-to-image subtasks.

Table 4 shows that directly using the model's posterior (bootstrapping) to assign alignments to a static subset of inputs leads to a degradation of learned representations and decreased performance compared to the baseline. Here the teacher network is likely reinforcing its own mistakes without penalty from the loss, leading to partial representation collapse for the unaligned examples. Utilizing the swapped prediction strategy restores zero-shot performance to slightly above baseline suggesting effective method for addressing this issue; however, retrieval performance is still lower than the baseline. By incorporating a dynamic partitioning strategy, our method exceeds baseline performance.

4.9. Qualitative Analysis

a. Distribution of Similarity Scores: Our method is largely motivated by the observation that previous VLP approaches neglect potential semantic similarity between negative samples and that accounting for this phenomenon can improve learned representations. In Figure 5, we plot the distribution of similarity scores for both positive and negative samples drawn from the COCO testing set. The left subplot reveals that our method consistently yields larger similarity scores for positive pairs compared to its CLIP counterpart and the OpenAI pretrained CLIP. Interestingly, the histogram of negative similarity scores shows that our method also assigns higher similarity scores to negative pairs. While it may seem counter intuitive to assign greater similarity scores to negative samples, we argue that doing so is the very reason our method captures greater similarity between positive pairs. By allowing some degree of alignment between the right set of negative examples, our method is able to minimize the inconsistencies between shared context of related positives and negatives. This in turn allows us to learn an overall more coherent representation space, resulting in increased robustness and downstream performance.

b. Visualizing Text-Image Retrievals: In Figure 6, we show the comparative list of top ten retrieved images for 5 example text queries. Overall, these retrievals suggest that

Pretraining Dataset	Method	Text-to-Image				Image-to-Text			
		R@1↑	R@5↑	R@10↑	MnR↓	R@1↑	R@5↑	R@10↑	MnR↓
COCO	CLIP	27.76	**57.34**	**70.70**	**23.10**	27.40	56.31	**68.65**	**20.11**
	Ours	**28.42**	57.14	68.86	26.11	**28.53**	**56.75**	68.10	22.01
CC3M	CLIP	12.50	29.76	40.92	91.04	9.88	24.86	35.30	106.94
	Ours	**16.98**	**37.12**	**48.28**	**63.80**	**13.19**	**31.54**	**43.00**	**72.92**
CC12M	CLIP	19.64	40.66	51.72	55.23	17.63	39.67	50.77	55.29
	Ours	**22.94**	**46.60**	**57.82**	**46.24**	**22.79**	**45.95**	**56.81**	**43.41**
CC3M	CLIP*	31.30	59.54	71.80	21.15	29.18	58.66	70.23	**17.02**
	Ours*	**33.26**	**66.70**	**76.92**	**18.77**	**32.20**	**60.14**	**71.96**	17.05
CC12M	CLIP*	35.22	62.74	73.46	17.70	33.36	61.15	73.36	15.76
	Ours*	**38.66**	**66.74**	**77.10**	**13.35**	**38.15**	**65.85**	**77.02**	**12.54**

Table 3. **Image-Text Retrieval** – Zero-shot retrieval performance on the COCO 5K testing set as measured by recall at 1,5,10 (higher is better) and mean rank (lower is better). The last four rows (marked with *) report the zero-shot retrieval results of the same models further finetuned on the COCO captions training set.
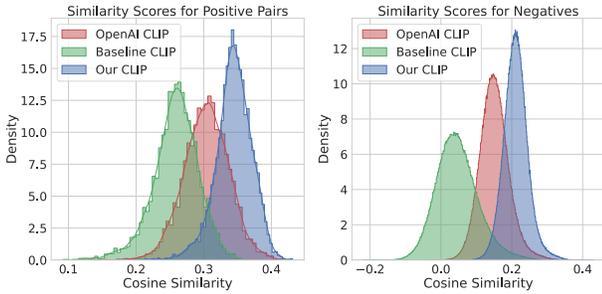

Figure 5. **Similarity Scores** – Similarity scores distribution for positive and negative pairs in joint cross-modal space after training with original CLIP loss and our proposed loss is provided. Compared to our baseline (COCO) pretrained CLIP and OpenAI's pretrained CLIP model, our method yields similarity scores with higher mean and lower variance for positive pairs. While COCO pretrained CLIP model concentrates negatives' similarity scores around zero, our method concentrates them at higher levels as it allows for some degree of semantic similarity between negatives.

our learned features more comprehensively capture all potential similarities between a text snippet and image. In contrast, baseline CLIP features tend to more narrowly focus on one specific commonality while neglecting others. For example, in the last row of Figure 6, CLIP strongly targets "living room" but misses the cat. Our approach on the other hand is successfully able to extract all the key aspects of the query including the cat, the chair and the living room. Similar trends are found in other examples.

5. Conclusions

We proposed a novel cross-modal contrastive learning framework with progressive self-distillation and soft image-text alignments. Our approach distills its own knowledge to dynamically generate soft-alignment targets for a subset of samples in every minibatch, which enables it to efficiently learn robust representations from noisy data. Exten-
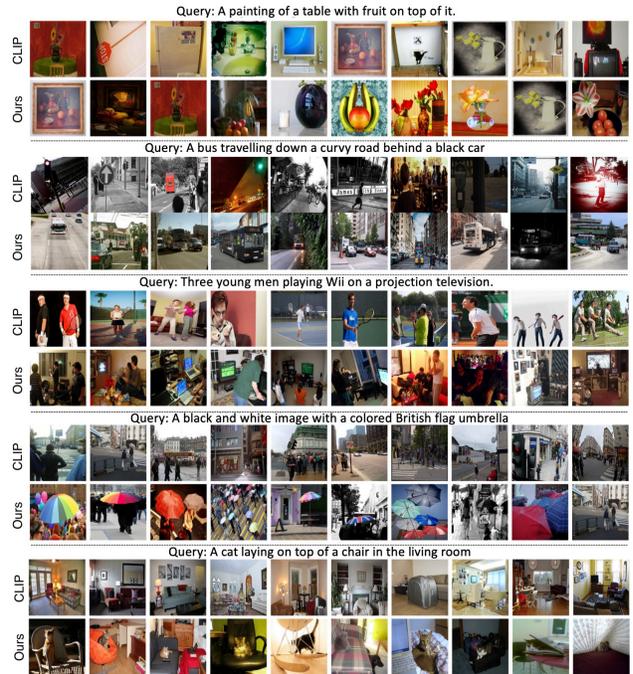

Figure 6. **Example Text-Image Retrievals** – Given a text query, we display the top ten most semantically related images (ranked left to right) retrieved by CLIP and our method. Compared to CLIP, our method continues to retrieve images that more holistically match the text description, even after the ground truth image has appeared in the ranking.

sive evaluations across 14 benchmark datasets showed that our method consistently outperforms its CLIP counterpart in multiple settings. Moreover, our method provides better effective robustness to natural distribution shifts compared to existing state-of-the-art methods. Going forward, we plan to further improve the efficiency of our approach by investigating data redundancy, network architectures and optimization algorithms during pretraining, so that we can generalize to even larger scale with fewer resources.

References

[1] Görkem Algan and Ilkay Ulusoy. Metalabelnet: Learning to generate soft-labels from noisy-labels. *arXiv preprint arXiv:2103.10869*, 2021. 1, 2

[2] Sanjeev Arora, Hrishikesh Khandeparkar, Mikhail Khodak, Orestis Plevrakis, and Nikunj Saunshi. A theoretical analysis of contrastive unsupervised representation learning. *arXiv preprint arXiv:1902.09229*, 2019. 2

[3] Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. 5, 6

[4] Thomas Berg, Jiongxin Liu, Seung Woo Lee, Michelle L. Alexander, David W. Jacobs, and Peter N. Belhumeur. Birdsnap: Large-scale fine-grained visual categorization of birds. In *Proc. Conf. Computer Vision and Pattern Recognition (CVPR)*, June 2014. 7

[5] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *European Conference on Computer Vision*, 2014. 7

[6] Cristian Buciluǎ, Rich Caruana, and Alexandru Niculescu-Mizil. Model compression. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 535–541, 2006. 2

[7] 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. 2

[8] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *arXiv preprint arXiv:2006.09882*, 2020. 2

[9] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pretraining to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3558–3568, 2021. 1, 2, 5, 6

[10] Guobin Chen, Wongun Choi, Xiang Yu, Tony Han, and Manmohan Chandraker. Learning efficient object detection models with knowledge distillation. *Advances in neural information processing systems*, 30, 2017. 2

[11] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning (ICML)*, 2020. 2

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Ieee, 2009. 5, 6, 7

[13] Aditya Deshpande, Jason Rock, and David Forsyth. Learning large-scale automatic image colorization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 567–575, 2015. 2

[14] Qianggang Ding, Sifan Wu, Hao Sun, Jiadong Guo, and Shu-Tao Xia. Adaptive regularization of labels. *arXiv preprint arXiv:1908.05474*, 2019. 2

[15] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE international conference on computer vision*, pages 1422–1430, 2015. 2

[16] Alexey Dosovitskiy, Philipp Fischer, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. Discriminative unsupervised feature learning with exemplar convolutional neural networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(9):1734–1747, 2015. 2

[17] Alexey Dosovitskiy, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. Discriminative unsupervised feature learning with convolutional neural networks. *Advances in neural information processing systems*, 27:766–774, 2014. 2

[18] Li Fei-Fei, Rob Fergus, and Pietro Perona. One-shot learning of object categories. *IEEE transactions on pattern analysis and machine intelligence*, 28(4):594–611, 2006. 6

[19] Andreas Fürst, Elisabeth Rumetshofer, Viet Tran, Hubert Ramsauer, Fei Tang, Johannes Lehner, David Kreil, Michael Kopp, Günter Klambauer, Angela Bitto-Nemling, et al. Cloob: Modern hopfield networks with infoloob outperform clip. *arXiv preprint arXiv:2110.11316*, 2021. 1, 2

[20] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018. 2

[21] Sangchul Hahn and Heeyoul Choi. Self-knowledge distillation in natural language processing. *arXiv preprint arXiv:1908.01851*, 2019. 2

[22] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Coteaching: Robust training of deep neural networks with extremely noisy labels. *arXiv preprint arXiv:1804.06872*, 2018. 2

[23] Tengda Han, Weidi Xie, and Andrew Zisserman. Video representation learning by dense predictive coding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 2

[24] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2, 4

[25] 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. *ICCV*, 2021. 1, 5, 6

[26] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples.(2019). *arXiv preprint cs.LG/1907.07174*, 2019. 5, 6

[27] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 2, 4

[28] Gabriel Ilharco, Mitchell Wortsman, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, July 2021. If you use this software, please cite it as below. 5

[29] J Yu Jason, Adam W Harley, and Konstantinos G Derpanis. Back to basics: Unsupervised learning of optical flow via brightness constancy and motion smoothness. In *European Conference on Computer Vision*, pages 3–10. Springer, 2016. 2

[30] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. *arXiv preprint arXiv:2102.05918*, 2021. 1

[31] Longlong Jing and Yingli Tian. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 2020. 2

[32] Dahun Kim, Donghyeon Cho, and In So Kweon. Self-supervised video representation learning with space-time cubic puzzles. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8545–8552, 2019. 2

[33] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[34] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 6

[35] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Learning representations for automatic colorization. In *European conference on computer vision*, pages 577–593. Springer, 2016. 2

[36] Hsin-Ying Lee, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Unsupervised representation learning by sorting sequences. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 667–676, 2017. 2

[37] Junnan Li, Yongkang Wong, Qi Zhao, and Mohan S Kankanhalli. Learning to learn from noisy labeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5051–5059, 2019. 2

[38] Tianhong Li, Jianguo Li, Zhuang Liu, and Changshui Zhang. Few sample knowledge distillation for efficient network compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14639–14647, 2020. 2, 4

[39] 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. *arXiv preprint arXiv:2110.05208*, 2021. 1, 2

[40] Yuncheng Li, Jianchao Yang, Yale Song, Liangliang Cao, Jiebo Luo, and Li-Jia Li. Learning from noisy labels with distillation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1910–1918, 2017. 2

[41] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 5

[42] Yuanze Lin, Xun Guo, and Yan Lu. Self-supervised video representation learning with meta-contrastive network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8239–8249, 2021. 4

[43] Haoliang Liu, Tan Yu, and Ping Li. Inflate and shrink: Enriching and reducing interactions for fast text-image retrieval. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9796–9809, 2021. 2

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

[45] Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, et al. Mixed precision training. *arXiv preprint arXiv:1710.03740*, 2017. 5

[46] John P Miller, Rohan Taori, Aditi Raghunathan, Shiori Sagawa, Pang Wei Koh, Vaishaal Shankar, Percy Liang, Yair Carmon, and Ludwig Schmidt. Accuracy on the line: On the strong correlation between out-of-distribution and in-distribution generalization. In *International Conference on Machine Learning*, pages 7721–7735. PMLR, 2021. 6

[47] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6707–6717, 2020. 2

[48] Ishan Misra, C Lawrence Zitnick, and Martial Hebert. Shuffle and learn: unsupervised learning using temporal order verification. In *European Conference on Computer Vision*, pages 527–544. Springer, 2016. 2

[49] Hossein Mobahi, Ronan Collobert, and Jason Weston. Deep learning from temporal coherence in video. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 737–744, 2009. 2

[50] Pedro Morgado, Ishan Misra, and Nuno Vasconcelos. Robust audio-visual instance discrimination. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12934–12945, 2021. 2, 3

[51] Pedro Morgado, Nuno Vasconcelos, and Ishan Misra. Audio-visual instance discrimination with cross-modal agreement. *CoRR*, abs/2004.12943, 2020. 2

[52] Pedro Morgado, Nuno Vasconcelos, and Ishan Misra. Audio-visual instance discrimination with cross-modal agreement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12475–12486, 2021. 1

[53] Rafael Müller, Simon Kornblith, and Geoffrey Hinton. When does label smoothing help? *arXiv preprint arXiv:1906.02629*, 2019. 2

[54] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European conference on computer vision*, pages 69–84. Springer, 2016. 2

[55] Curtis Northcutt, Lu Jiang, and Isaac Chuang. Confident learning: Estimating uncertainty in dataset labels. *Journal of Artificial Intelligence Research*, 70:1373–1411, 2021. 2

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

[57] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3498–3505. IEEE, 2012. 7

[58] Deepak Pathak, Ross Girshick, Piotr Dollár, Trevor Darrell, and Bharath Hariharan. Learning features by watching objects move. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2701–2710, 2017. 2

[59] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544, 2016. 2

[60] Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. Making deep neural networks robust to label noise: A loss correction approach. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1944–1952, 2017. 2

[61] Gabriel Pereyra, George Tucker, Jan Chorowski, Łukasz Kaiser, and Geoffrey Hinton. Regularizing neural networks by penalizing confident output distributions. *arXiv preprint arXiv:1701.06548*, 2017. 2

[62] AJ Piergiovanni, Anelia Angelova, and Michael S Ryoo. Evolving losses for unsupervised video representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 133–142, 2020. 2

[63] 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. *arXiv preprint arXiv:2103.00020*, 2021. 1, 2, 3, 5, 6

[64] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, pages 5389–5400. PMLR, 2019. 5, 6

[65] Scott Reed, Honglak Lee, Dragomir Anguelov, Christian Szegedy, Dumitru Erhan, and Andrew Rabinovich. Training deep neural networks on noisy labels with bootstrapping. *arXiv preprint arXiv:1412.6596*, 2014. 2, 3

[66] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014. 2

[67] 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 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018. 1, 2, 5, 6

[68] Sheng Shen, Liunian Harold Li, Hao Tan, Mohit Bansal, Anna Rohrbach, Kai-Wei Chang, Zhewei Yao, and Kurt Keutzer. How much can clip benefit vision-and-language tasks? *arXiv preprint arXiv:2107.06383*, 2021. 1, 2

[69] Jun Shu, Qi Xie, Lixuan Yi, Qian Zhao, Sanping Zhou, Zongben Xu, and Deyu Meng. Meta-weight-net: Learning an explicit mapping for sample weighting. *arXiv preprint arXiv:1902.07379*, 2019. 2

[70] Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 1, 2, 6

[71] 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*, 59(2):64–73, 2016. 1, 2

[72] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *European Conference on Computer Vision (ECCV)*, 2020. 2

[73] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 776–794. Springer, 2020. 2

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

[75] Jue Wang, Haofan Wang, Jincan Deng, Weijia Wu, and Debing Zhang. Efficientclip: Efficient cross-modal pretraining by ensemble confident learning and language modeling. *arXiv preprint arXiv:2109.04699*, 2021. 1, 2

[76] Yisen Wang, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, and James Bailey. Symmetric cross entropy for robust learning with noisy labels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 322–330, 2019. 2

[77] Zhirong Wu, Yuanjun Xiong, X Yu Stella, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[78] Ting-Bing Xu and Cheng-Lin Liu. Data-distortion guided self-distillation for deep neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 5565–5572, 2019. 2

[79] Zhichao Yin and Jianping Shi. Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1983–1992, 2018. 2

[80] Sukmin Yun, Jongjin Park, Kimin Lee, and Jinwoo Shin. Regularizing class-wise predictions via self-knowledge distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13876–13885, 2020. 2

[81] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European conference on computer vision*, pages 649–666. Springer, 2016. 2

[82] Richard Zhang, Phillip Isola, and Alexei A Efros. Split-brain autoencoders: Unsupervised learning by cross-channel prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1058–1067, 2017. 2

[83] Zhilu Zhang and Mert R Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. In *32nd Conference on Neural Information Processing Systems (NeurIPS)*, 2018. 2

[84] Zizhao Zhang, Han Zhang, Sercan O Arik, Honglak Lee, and Tomas Pfister. Distilling effective supervision from severe label noise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9294–9303, 2020. 2

[85] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017. 5, 6

[86] Chengxu Zhuang, Tianwei She, Alex Andonian, Max Sobol Mark, and Daniel Yamins. Unsupervised learning from video with deep neural embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9563–9572, 2020. 2

[87] Chengxu Zhuang, Alex Lin Zhai, and Daniel Yamins. Local aggregation for unsupervised learning of visual embeddings. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6002–6012, 2019. 2

[88] Mohammadreza Zolfaghari, Yi Zhu, Peter Gehler, and Thomas Brox. Crossclr: Cross-modal contrastive learning for multi-modal video representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1450–1459, 2021. 2