CyCLIP: Cyclic Contrastive Language-Image Pretraining

Anonymous Author(s)
Affiliation
Address
email

Abstract

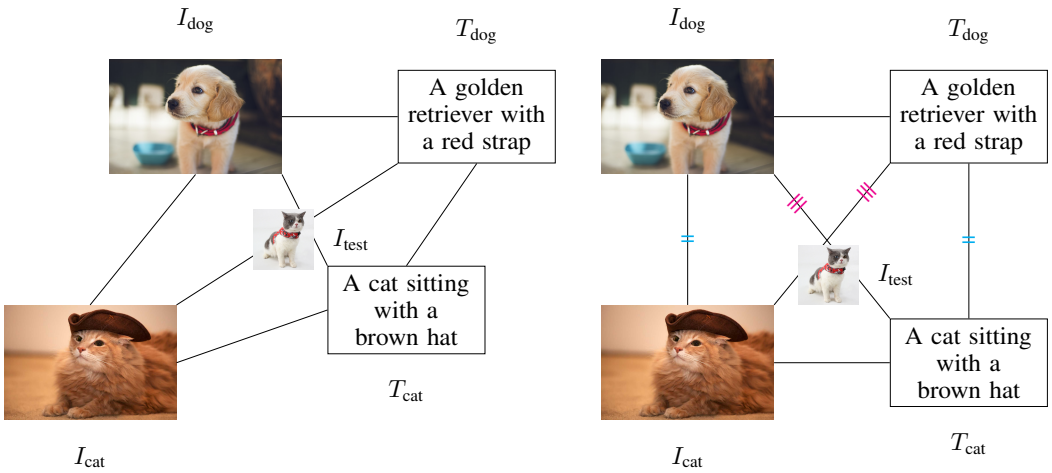
Recent advances in contrastive representation learning over paired image-text data have led to models such as CLIP [43] that achieve state-of-the-art performance for zero-shot classification and distributional robustness. Such models typically require joint reasoning in the image and text representation spaces for downstream inference tasks. Contrary to prior beliefs, we demonstrate that the image and text representations learned via a standard contrastive objective are not interchangeable and can lead to inconsistent downstream predictions. To mitigate this issue, we formalize *consistency* and propose CyCLIP, a framework for contrastive representation learning that explicitly optimizes for the learned representations to be *geometrically consistent* in the image and text space. In particular, we show that consistent representations can be learned by explicitly symmetrizing (a) the similarity between the two mismatched image-text pairs (cross-modal consistency); and (b) the similarity between the image-image pair and the text-text pair (in-modal consistency). Empirically, we show that the improved consistency in CyCLIP translates to significant gains over CLIP, with gains ranging from $10\% - 24\%$ for zero-shot classification accuracy on standard benchmarks (CIFAR-10, CIFAR-100, ImageNet1K) and $10\% - 27\%$ for robustness to various natural distribution shifts.

1 Introduction

The ability to learn general-purpose representations from diverse data modalities is a long-standing goal of artificial intelligence (AI) [4, 31]. In this regard, recent instantiations such as CLIP [43], ALIGN [28], and BASIC [40] have scaled up vision-language contrastive pretraining to jointly learn image and text embeddings, by exploiting an enormous amount of paired image-text data on the web. Post pretraining, these embeddings exhibit impressive zero-shot classification performance [13] and robustness to natural distribution shift benchmarks [47, 56, 23, 25]. Recently, these embeddings have been extended to text-guided generation of natural images [46, 12, 36, 45] and transferred to modalities such as 3-D shapes [49] by emphasizing the interchangeability of the image and text embeddings.

In the context of vision-language pretraining, the standard contrastive learning objective aims to maximize the similarity between matched image-text pairs (“positives") against all the mismatched image-text pairs (“negatives") [44, 7, 39, 21]. While such an objective aligns the true image-text pairs, it poses no constraints on the overall geometry of all data pairs, including the mismatched pairs and pairs within the same modality. In Figure 1 (a), we illustrate this effect where matched image-text pairs, $(I_{\text{dog}}, T_{\text{dog}})$ and $(I_{\text{cat}}, T_{\text{cat}})$, get close to each other but the overall geometry of pairwise distances can be highly irregular (see e.g., $(I_{\text{dog}}, T_{\text{cat}})$ and $(I_{\text{cat}}, T_{\text{dog}})$). If we use such representations for downstream inference, such irregularities can translate into inconsistent reasoning in the image and text spaces. For example, CLIP designs proxy captions for class labels and uses the

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.



(a) CLIP (b) CyCLIP

Figure 1: An illustration of the planar geometry of the learned representations of image-text pairs by (a) CLIP and (b) CyCLIP. The edges indicate the distance between the representations i.e., $d(e_1, e_2) = 1 - \langle e_1, e_2 \rangle$, where $\langle \cdot, \cdot \rangle$ is the inner product. CyCLIP is cyclic consistent between image-text pairs as the in-modal distances, $d(T_{\text{cat}}, T_{\text{dog}}) \sim d(I_{\text{cat}}, I_{\text{dog}})$, and the cross-modal distances, $d(T_{\text{cat}}, I_{\text{dog}}) \sim d(I_{\text{cat}}, T_{\text{dog}})$, are similar to each other unlike CLIP. Due to explicit consistency constraints, the test image of a cat is classified as a cat in the image as well as the text space.

most similar class caption to perform zero-shot classification for images; using the default captions in Figure 1 (a), this would imply that a test image I_{test} gets classified as a dog in the image space even when a simple nearest neighbor classifier in the text space would correctly infer the label to be a cat.

To mitigate these challenges, we propose **Cy**clic **C**ontrastive **L**anguage-**I**mage **P**retraining (CyCLIP), a framework that imposes additional geometric structure on the learned representations. Specifically, given two image-text pairs, we augment the contrastive learning objective with two symmetrization terms. The first term provides for in-modal consistency by encouraging the distance between the two image embeddings to be close to the distance between the corresponding text embeddings. The second term for the cross-modal consistency that encourages the distance between the image and text embedding from the first and second pairs respectively to be close to the distance between the text and image embeddings from the first and second pairs respectively. As shown in Figure 1 (b), if representations of any two image-text pairs, $(I_{\text{dog}}, T_{\text{dog}})$ and $(I_{\text{cat}}, T_{\text{cat}})$ exactly satisfy both forms of cyclic consistency, then we can guarantee that any test image I_{test} respects the ordering of distances in both image and text spaces (i.e., if $d(I_{\text{test}}, I_{\text{dog}}) > d(I_{\text{test}}, I_{\text{cat}})$, then $d(I_{\text{test}}, T_{\text{dog}}) > d(I_{test}, T_{\text{cat}})$).

Empirically, we demonstrate that the improved consistency in CyCLIP translates to improvements over CLIP. In all cases, we pre-train our models on the Conceptual Captions 3M dataset[51]. On zero-shot classification, we observe that CyCLIP improves over CLIP by 10.2% on ImageNet1K, 10.6% on CIFAR-10 and 23.9% on CIFAR-100 respectively. Further, CyCLIP outperforms CLIP with an average relative gain of $+17\%$ on ImageNet natural distribution shift benchmarks. We further analyze the improved performance of CyCLIP and find that the additional geometric structure in the representation space better captures the coarse and fine-grained concept hierarchies of datasets.

Our contributions are as follows:

1. We analyze contrastive learning for representation learning jointly over image and text modalities. We identify a critical shortcoming in the geometry of the learned representation space that can lead to inconsistent predictions in image and text domains.
2. We propose CyCLIP, a simple and effective framework for contrastive representation learning with two additional cycle consistency constraints for mitigating the above issue.
3. We demonstrate that CyCLIP achieves significant empirical improvements over CLIP on zero-shot classification and robustness benchmarks. We further explain these improvements by analyzing the impact of consistency on the hierarchical structure of datasets.

2 Cycle Consistent Representation Learning

2.1 Preliminaries

We are interested in using text supervision to learn general-purpose visual representations that can be generalized to downstream predictive tasks. To this end, there have been several recent advances in language-image pretraining concerning model architectures, training objectives, and sources of supervision. Our work is most closely related to Contrastive Language-Image Pretraining (CLIP) [43] which combines many such advances in a highly scalable and generalizable learning framework.

CLIP is trained on millions of images with their captions scraped from the web. Formally, we consider a dataset $S \subset \mathcal{I} \times \mathcal{T}$ consisting of pairs (I_j, T_j) where I_j is a raw image and T_j is a text caption. We use $\mathcal{I}$ and $\mathcal{T}$ to denote the domain of images and text, respectively. The CLIP architecture consists of 3 components: (i) an image encoder network, $f_I : \mathcal{I} \mapsto \mathbb{R}^d$, to encode the raw image into an embedding vector of dimension d, (ii) a text encoder network, $f_T : \mathcal{T} \mapsto \mathbb{R}^d$, to encode the raw text into an embedding vector of dimension d, (iii) a contrastive objective that pulls the embeddings of paired image-caption pairs together while pushing apart embeddings of unmatched pairs.

Formally, during training, consider a batch of N image-captions pairs, $\{I_j, T_j\}_{j=1}^N$, where I_j and T_j represent the raw image and text pair, respectively. The image embedding $I_j^e \in \mathbb{R}^d$ and text embedding $T_j^e \in \mathbb{R}^d$ are obtained by passing I_j and T_j through the image encoder f_I and text encoder f_T, respectively; i.e. $I_j^e = f_I(I_j)$ and $T_j^e = f_T(T_j)$. Further, we assume they are normalized to have unit ℓ_2-norm. The contrastive objective in CLIP aims to align the image and text representations by minimizing the loss function $\mathcal{L}_{\text{CLIP}}$ shown below:

$$\mathcal{L}_{\text{CLIP}} = -\frac{1}{2N}\sum_{j=1}^{N} \log \underbrace{\left[\frac{\exp\left(\langle I_j^e, T_j^e\rangle/\tau\right)}{\sum_{k=1}^{N}\exp\left(\langle I_j^e, T_k^e\rangle/\tau\right)}\right]}_{\text{Contrasting images with the texts}} - \frac{1}{2N}\sum_{k=1}^{N} \log \underbrace{\left[\frac{\exp\left(\langle I_k^e, T_k^e\rangle/\tau\right)}{\sum_{j=1}^{N}\exp\left(\langle I_j^e, T_k^e\rangle/\tau\right)}\right]}_{\text{Contrasting texts with the images}} \tag{1}$$

where $\langle \cdot, \cdot \rangle$ represents the inner product, and τ is a trainable temperature parameter. CLIP and its variants can be used to perform zero-shot image classification, i.e., classifying test images into categories not seen at training time. We first transform each category into a suitable caption (e.g., the airplane category in CIFAR-10 can be expressed as 'a photo of an airplane'). Then, the similarity of the test image to each caption is computed (e.g., cosine distance), and the model predicts the category for which the image-caption similarity is the highest.

2.2 Inconsistent Representation Learning in CLIP

As illustrated in Figure 1 (a), the standard contrastive objective in CLIP can learn image-text representations such that the predicted labels for the test image are different in the image and text spaces. Here, we reason about such inconsistencies more formally in the context of downstream classification. As discussed above, we can predict a label in the text embedding space (zero-shot setting) by selecting the label that is closest to the test image (P_T). Additionally, for classification in the image embedding space, if we had access to a labeled training set, then one natural way to infer the predicted label (P_I^k) of a test image I_{test} is by taking a majority vote from the true labels associated with the k-nearest training images. Formally, we define a consistency score that measures the synchrony between the predicted labels in the image and text spaces as:

$$\text{Consistency Score}_k = \frac{1}{N}\sum_{j=1}^{N} \mathbb{1}\left[P_I^k(I_j) = P_T(I_j)\right] \tag{2}$$

where N is the number of test images. In our experiments (discussed in detail in §3), we found the CLIP's consistency score ($k = 1$) to be 44%, 16%, and 16% on the standard benchmarks CIFAR-10, CIFAR-100, and ImageNet1K, respectively, showing a very high degree of disagreement in the image and text spaces. In the following section, we describe our approach to alleviate the inconsistent inference problem and quantitatively show that our solution improves the consistency score in §4.1.

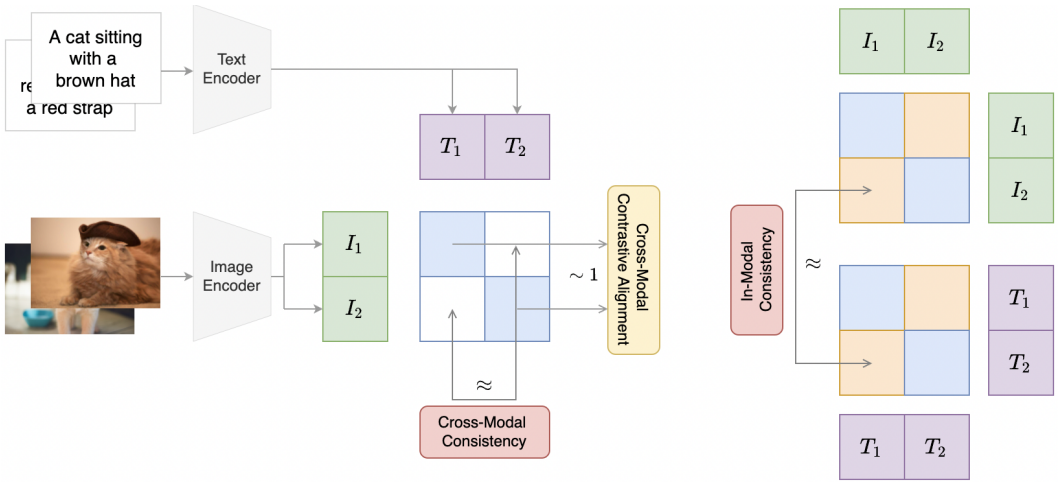


Figure 2: Illustrative overview for CyCLIP ($N = 2$). It consists of 3 major components: (a) cross-modal contrastive alignment, (b) cross-modal consistency, and (c) in-modal consistency. Only (a) is present in CLIP, whereas our proposed regularizers in (b) and (c) mitigate inconsistency.

2.3 Cycle Consistent Representation Learning via CyCLIP

We showed that the visual representations learned by CLIP could be inconsistent when used for inference in the image and text spaces. To mitigate this problem, we propose CyCLIP, a learning framework that builds upon CLIP by augmenting the contrastive loss in Eq. 1 with additional geometric consistency regularizers. The intuition follows directly from Figure 1 (b), where we showed that inconsistency in the image and text spaces could be eliminated if we symmetrize the similarity between the two mismatched image-text pairs and the similarity between the image-image pair and the text-text pair. We formalize this intuition with two consistency regularizers.

(1) The **cross-modal consistency** regularizer reduces the gap in the similarity scores between the embeddings of all the mismatched image-text pairs in a batch, two at a time:

$$\mathcal{L}_{\text{C-Cyclic}} = \frac{1}{N}\sum_{j=1}^{N}\sum_{k=1}^{N}\left(\langle I_j^e, T_k^e\rangle - \langle I_k^e, T_j^e\rangle\right)^2. \tag{3}$$

(2) The **in-modal consistency** regularizer reduces the gap in the similarity scores between the embeddings of all combinations of image pairs and their corresponding text pairs in a batch:

$$\mathcal{L}_{\text{I-Cyclic}} = \frac{1}{N}\sum_{j=1}^{N}\sum_{k=1}^{N}\left(\langle I_j^e, I_k^e\rangle - \langle T_k^e, T_j^e\rangle\right)^2. \tag{4}$$

Hence, our overall loss for CyCLIP is given as:

$$\mathcal{L}_{\text{CyCLIP}} = \mathcal{L}_{\text{CLIP}} + \lambda_1\mathcal{L}_{\text{I-Cyclic}} + \lambda_2\mathcal{L}_{\text{C-Cyclic}} \tag{5}$$

where $\lambda_1 > 0$ and $\lambda_2 > 0$ are hyperparameters controlling the importance of the in-modal and cross-modal cyclic consistency regularizers relative to the contrastive loss in CLIP. We can also characterize the effect of the regularizers in terms of symmetrizing the in-modal and cross-modal similarity matrices, as illustrated in Figure 2. Note that the optimal solution to the contrastive loss formulation would push the similarity between the normalized embeddings of the matched pairs towards 1 while forcing all other pairs of similarities to 0, thereby also symmetrizing the cross-modal similarity matrix and minimizing the cross-modal consistency loss. However, this idealized scenario does not occur in practice, and we find that explicit regularization via cycle-consistency in CyCLIP facilitates improved learning, as we show in our experiments.

3 Experiments

Setup: We use Conceptual Captions 3M [51] (CC3M) image-caption pairs as the source of multimodal pretraining data for all our models. Note while this dataset is smaller than the custom dataset (400 million pairs) used in the original work on CLIP [43], it is suitable for our available data and compute and has been used for benchmark evaluations in many subsequent works on language-image pretraining [5, 32, 35, 55]. Following prior work [43], our CLIP models use ResNet-50 as the image encoder and a transformer architecture as the text encoder. Further, we train our models from scratch for 64 epochs on 4 V100 GPUs with a batch size of 128 and an initial learning rate of 0.0005 with cosine scheduling and 10000 warmup steps. The dimension of the image and text embeddings is 1024. For CyCLIP, we use $\lambda_1 = 0.25$ and $\lambda_2 = 0.25$ across all our experiments.

3.1 Zero-Shot Transfer

We compare the zero-shot performance of CLIP and CyCLIP on standard image classification datasets: CIFAR-10, CIFAR-100 [30], and ImageNet1K [48]. We follow the evaluation strategy suggested by [43] for zero-shot classification using prompt engineering. For each dataset, we use the names of the classes to form a set of natural sentences such as 'a photo of the {class name}', 'a sketch of the {class name}' and more. These are passed through the text encoder to get a set of text embeddings for that class. This set of text embeddings are ℓ_2-normalized, averaged, and further ℓ_2-normalized to obtain a single text embedding for that class. For a given image, the image embedding is obtained as described in §2. The class whose text embedding (as described above) is closest to the test image is taken to be the predicted label. The zero-shot performance of the models is presented in Table 1.

Table 1: Zero-shot TopK classification accuracy (%) where K$\in \{1, 3, 5\}$

	CIFAR-10			CIFAR-100			ImageNet1K		
	Top1	Top3	Top5	Top1	Top3	Top5	Top1	Top3	Top5
CLIP	46.54	78.22	91.16	18.69	34.72	43.97	20.03	33.04	39.35
CyCLIP	**51.45**	**79.57**	**91.80**	**23.15**	**41.46**	**50.66**	**22.08**	**35.98**	**42.30**
%Gain	+10.6	+1.7	+0.7	+23.9	+19.4	+15.2	+10.2	+8.9	+7.5

We observe that the CyCLIP outperforms CLIP across all the datasets and on all TopK metrics, with gains in the range of 10% - 24% for K$= 1$. Our results on zero-shot transfer indicate the usefulness of having geometrical consistency for improved downstream performance of CLIP.

3.2 Robustness to Natural Distribution Shifts

One of the major successes of CLIP was its state-of-the-art performance on the natural distribution shift benchmarks. These benchmarks include images depicting sketches, cartoons, adversaries generated using attacks on trained ImageNet models. In Table 2, we evaluate the zero-shot classification accuracy of CyCLIP on four natural distribution shift benchmarks for the ImageNet dataset: ImageNetV2 [47], ImageNetSketch [56], ImageNet-A [26], and ImageNet-R [24].

Table 2: Zeroshot Classification on Natural Distribution Shifts (%)

	ImageNetV2			ImageNetSketch			ImageNet-A			ImageNet-R		
	Top1	Top3	Top5	Top1	Top3	Top5	Top1	Top3	Top5	Top1	Top3	Top5
CLIP	16.91	29.28	34.99	10.37	19.15	24.20	4.23	11.35	16.88	24.32	39.69	47.20
CyCLIP	**19.22**	**32.29**	**38.41**	**12.26**	**22.56**	**28.17**	**5.35**	**13.53**	**19.51**	**26.79**	**42.31**	**50.03**
%Gain	+13.7	+10.3	+9.8	+18.2	+17.8	+16.4	+26.5	+19.2	+15.6	+10.2	+6.6	+6.0

For most of the distribution shift benchmarks, both CLIP and CyCLIP undergo a significant reduction in their zero-shot performance compared to the original ImageNet1K dataset (last three columns in

Table 1). However, we observe that CyCLIP outperforms CLIP on all of the datasets considered in this experiment by a significant margin of improvement (10 - 27%). This result indicates that having cyclic consistency in the learned representations preserves the robustness on the traditional datasets.

3.3 Linear Probing

While the primary focus of CLIP and CyCLIP is zero-shot generalization, we can also assess if the benefits of our cyclic consistency constraints in mitigating inconsistency can be recovered with extra in-domain and in-modality supervision i.e., in the presence of in-distribution training samples from in-domain visual datasets. To this end, we conduct an additional experiments on linear probing where we fit a linear classifier on the representations learned by the visual encoder (ResNet-50) of CLIP and CyCLIP on a range of image classification datasets.

Table 3: Transfer CLIP and CyCLIP to 14 downstream visual datasets using linear probing. Our CyCLIP performs marginally better on 10 out of 14 datasets. For training ImageNet1K, we use a random subset of 50K images from its original training dataset.

	Caltech101	CIFAR-10	CIFAR-100	DTD	Aircraft	Flowers102	Food101	GTSRB	ImageNet1K	OxfordPets	SST2	StanfordCars	STL10	SVHN	Average
CLIP	79.98	**78.27**	54.85	61.44	**28.02**	**83.74**	54.47	69.45	35.47	**58.30**	53.10	20.53	89.76	47.64	58.22
CyCLIP	**80.88**	77.34	**55.74**	**64.47**	27.87	82.97	**54.95**	**71.70**	**36.69**	57.29	**54.04**	**22.72**	**90.42**	**48.16**	**58.95**

We present our results in Table 3. We find that both CLIP and CyCLIP can recover most of the performance lost due to inconsistency when provided extra in-domain and in-modality supervision, with CyCLIP marginally outperforming the CLIP on 10 out of 14 visual datasets.

4 Analysis

Previously, we demonstrated the gains of CyCLIP over CLIP on downstream tasks that involve joint reasoning over the image and text spaces. In the current section, we wish to better understand the relative behavior of the two models on a set of challenging tasks.

4.1 Consistency in Image and Text Spaces

We begin by quantitatively measuring the inconsistency problem illustrated in Figure 1. That is, we wish to evaluate to what extent are the predictions in the image-text space (zero-shot) consistent with the ones made purely within the image space, as measured by our consistency metric in Eq. 2.

Table 4: Consistency score (%) trend for CLIP and CyCLIP across standard benchmarks . Top-k consistency score implies the fraction of times, the zero-shot predicted label in the text space is identical to the k-Nearest Neighbor predicted label in the image space (using the training dataset).

	CIFAR-10				CIFAR-100				ImageNet1K			
	Top1	Top3	Top5	Top10	Top1	Top3	Top5	Top10	Top1	Top3	Top5	Top10
CLIP	44.60	46.04	47.06	48.45	16.21	17.28	18.42	19.36	16.34	17.42	18.58	19.78
CyCLIP	**48.81**	**50.89**	**52.30**	**53.71**	**20.43**	**21.96**	**23.18**	**24.31**	**19.20**	**20.31**	**21.95**	**23.94**
%Gain	+8.6	+9.5	+10.0	+9.8	+20.7	+21.3	+20.5	+20.4	+14.9	+14.2	+15.4	+17.4

Table 4 presents our results over standard benchmarks (CIFAR-10, CIFAR-100, ImageNet1K). The consistency score is calculated over 10K, 10K, and 50K testing images of the CIFAR-10, CIFAR-100 and ImageNet dataset respectively. We use 50K samples from the training set of each dataset for k-Nearest Neighbor prediction. CyCLIP is more consistent than CLIP across all the datasets as we

explicitly symmetrize the cross-modal and in-modal distances. Hence, the representations learned by CyCLIP can be better used interchangeably than CLIP.

4.2 Fine-grained and Coarse-grained Performance

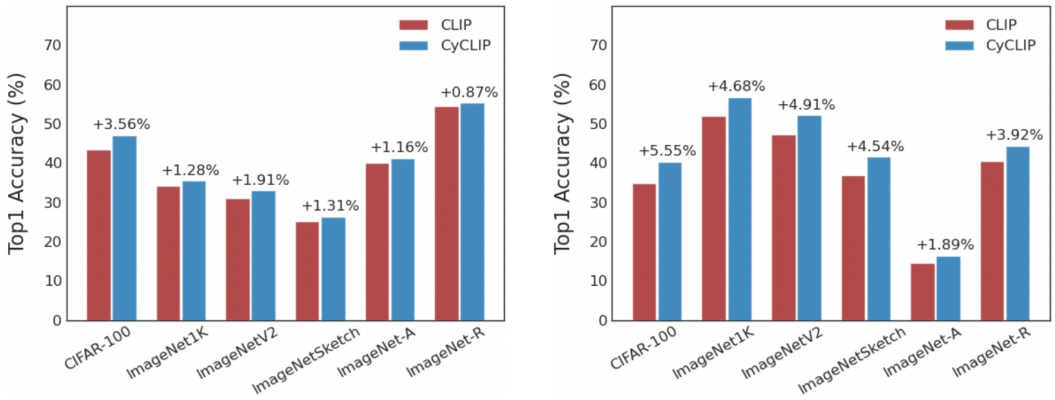
In §3.1, we observed that CyCLIP outperforms CLIP on zero-shot transfer across various datasets. We perform an error analysis investigating both models' coarse and fine-grained classification performance to understand the transfer phenomena better. Given a hierarchical class structure dataset, coarse-grained classification differentiates between high-level (parent) classes, i.e., zero-shot classification into aquatic mammals and fish. The fine-grained classification task focuses on differentiating low-level (child) classes, i.e., zero-shot classification into a dolphin, otter, and seal (subclasses of aquatic mammals). We perform this analysis on the CIFAR-100, ImageNet1K, ImageNetV2, ImageNetSketch, ImageNet-A, and ImageNet-R datasets.

Formally, we consider a test set of N image-subclass-superclass triplets, $\{I_j, C_j, P_j\}_{j=1}^{N}$, where I_j, C_j, P_j represent the image, the subclass (child) and superclass (parent) respectively. The image embedding $I_j^e \in \mathbb{R}^d$ is obtained as described in §2, and the subclass embedding $C_j^e \in \mathbb{R}^d$ and superclass embedding $P_j^e \in \mathbb{R}^d$ are obtained as described in §3.1. Let the total number of superclasses and subclasses in the dataset be n_{p} and n_{c} , respectively. Further, let F be a unique mapping from a subclass to the superclass, and G denote the inverse mapping from a superclass to the set of subclasses i.e. $\forall P \in \{1, \ldots, n_{\text{p}}\}$, $G(P) = \{C : F(C) = P \text{ and } C \in \{1, \ldots, n_{\text{c}}\}\}$. Under this setup, the fine-grained and coarse-grained accuracies are defined as:

$$\text{Fine-grained Accuracy} = \frac{1}{N} \sum_{j=1}^{N} \mathbb{1} \left[\underset{C \in G(P_j)}{\text{argmax}} \langle I_j^e, C \rangle = C_j \right] \tag{6}$$

$$\text{Coarse-grained Accuracy} = \frac{1}{N} \sum_{j=1}^{N} \mathbb{1} \left[\underset{C \in \{1, \ldots, n_{\text{c}}\}}{\text{argmax}} \langle I_j^e, C \rangle \in G\left(P_j\right) \right] \tag{7}$$

In Figure 3 we visualize how CLIP and CyCLIP compare with each other on the above metrics. The difference between the zero-shot performance of CyCLIP and CLIP is much more significant for coarse-grained classification than fine-grained classification across all the datasets. This observation indicates that concept-level knowledge is better captured in CyCLIP compared to CLIP. The drastic difference in the coarse-grained performance of CyCLIP and CLIP may be attributed to the rigid separation that the default cross-entropy loss in CLIP enforces between the positive pairs and negative pairs, which might degrade performance when some pairs in the negative batch belong to a similar entity. However, CyCLIP does not suffer from this problem as much because it poses cycle constraints on the overall geometry of all the data pairs rather than forcing a rigid separation.



(a) Fine-grained (b) Coarse-grained

Figure 3: The gap between the performances of CLIP and CyCLIP is much larger in coarse-grained scenario highlighting better entity-level knowledge representation in CyCLIP.

4.3 Alignment and Uniformity on the Unit Hypersphere

[57] argues that contrastive learning directly optimizes for (a) alignment (closeness) of the representations of the positive pairs and (b) uniformity (coverage) of the representation space on the unit hypersphere. We extend these properties for multimodal contrastive representation learning as:

$$\text{Alignment} = \frac{1}{N}\sum_{j=1}^{N}\langle I_j^e, T_j^e\rangle \qquad \text{Uniformity} = \log\left(\frac{1}{N(N-1)}\sum_{j=1}^{N}\sum_{k=1,j\neq k}^{N} e^{-\langle I_j^e, T_k^e\rangle}\right) \tag{8}$$

We desire our models to achieve high alignment and uniformity scores so that the image-text representations are close for the matched pairs and better spread over the unit hypersphere for different categories. We analyze the effect of cross-modal and in-modal consistency on the alignment and uniformity of the shared representations. For this, we train two ablated versions of CyCLIP, 1) C-CyCLIP with only cross-modal consistency component i.e. $\lambda_1 = 0, \lambda_2 = 0.5$, and 2) I-CyCLIP with only in-modal consistency component i.e. $\lambda_1 = 0.5, \lambda_2 = 0$ (in Eq. 5). We design proxy captions for classes as discussed in §3.1 to act as text embeddings. We present the results in Table 5.

Table 5: Alignment and Uniformity values for CLIP and Cyclic CLIP models. We abbreviate Alignment by A, Uniformity by U, and Zero-shot Top1 classification accuracy (%) by ZS-Top1.

Model	CIFAR-10			CIFAR-100			ImageNet1K		
	A	U	ZS-Top1	A	U	ZS-Top1	A	U	ZS-Top1
CLIP	0.36	-0.27	46.54	0.36	-0.25	18.69	0.39	-0.18	20.03
CyCLIP	0.36	-0.34	51.45	0.37	-0.33	23.15	0.38	-0.32	**22.08**
I-CyCLIP	**0.60**	-0.57	50.97	**0.60**	-0.57	22.35	**0.61**	-0.55	21.21
C-CyCLIP	0.05	**-0.02**	**55.52**	0.06	**-0.02**	**25.49**	0.07	**-0.02**	21.73

We observe that I-CyCLIP learns representations that are better aligned in the representation space; however, they do not cover the hypersphere uniformly. The representations learned by C-CyCLIP are more uniformly spread but poorly aligned compared to I-CyCLIP. In this light, the components of CyCLIP can be seen to encourage a balance of good alignment and uniformity. Further, we find that CLIP is more uniform than CyCLIP in all datasets, but contrary to prior beliefs, this does not translate to improved downstream performance. C-CyCLIP has the best downstream zero-shot performance for CIFAR-10 and CIFAR-100 despite its poor alignment score. Further, all 3 variants of CyCLIP outperform CLIP on all 3 datasets, with CyCLIP performing the best on ImageNet1K.

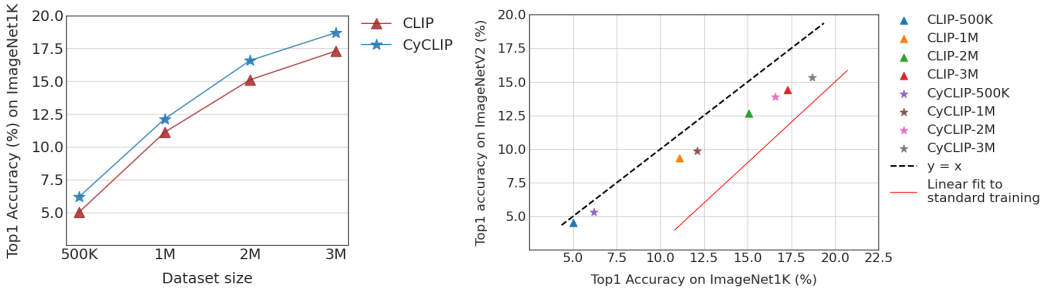
4.4 CyCLIP preserves the Effective Robustness of CLIP

[34] shows that there is a strong correlation between the in-distribution and out-of-distribution generalization of the models trained on ImageNet1K, as illustrated by the linear fit (red) in Figure 4. Ideally, any model that does not undergo distribution shift would fall on the $y = x$ trendline (black). For other models, the deviations of the models from this ideal fit indicate their effective robustness. Previously, [44] showed that the zero-shot CLIP classifier trained on 400M image-text pairs improves effective robustness significantly compared to prior approaches to robustness. Subsequently, [27] demonstrated that CLIP models trained at small scales also exhibit high effective robustness that allows them to be used as a proxy to study the robustness properties of CLIP.

We evaluate the effect of cyclic consistency on effective robustness. We trained 4 CLIP and CyCLIP models, varying the training dataset sizes from 500K to 3M image-text pairs from the CC3M dataset. In Figure 4, (a) we observe that for all training data sizes, CyCLIP shows a significant improvement over CLIP, showcasing its effectiveness in a diverse set of data regimes. Further, Figure 4 (b) shows that CyCLIP lies way above the baseline trend and preserves the effective robustness of CLIP.

5 Related Work

Our work fits into the broader theme of unsupervised pretraining with multiple modalities and has been successfully applied for learning representations of modalities such as images, text, and speech [2, 15, 1, 58, 42]. Similar to the unimodal setting, two predominant approaches for multimodal pretraining are contrastive and generative, as described below.



(a) (b)

Figure 4: Effect of varying the training dataset size on (a) Classification accuracy on ImageNet1K and (b) Effective Robustness on ImageNetV2.

Contrastive Representation Learning: Contrastive learning was originally proposed for self-supervised representation learning in the unimodal context where the embeddings of a sample are brought closer to an augmented version of the sample. In contrast, the embeddings are pushed away for other samples, and their augmentation [11, 50, 38, 54, 20, 7, 16, 39, 65, 22, 18]. [62] and [3] impose additional constraints to remove redundancies and prevent dimensional collapse in the visual representations. Recently, contrastive learning has also been used to learn robust representations of the multimodal data [61, 46]. Many works use additional losses to imbibe extra supervisory multimodal knowledge during the training process [53, 64, 63, 14, 33]. In this work, we focus on having cyclic consistency in addition to the contrastive loss to learn more robust image-text representations.

Contrastive Language-Image Pretraining: CLIP [43], ALIGN [28] and BASIC [40] have enjoyed great success in extending contrastive learning to paired image-text data, with impressive zero-shot classification and robustness performance. These works have been further extended recently to include visual self-supervision [35], additional nearest neighbor supervision [32], and utilization of unpaired data [55]. Our work complements much of this literature as it identifies consistency regularizers that can be augmented to the learning objective of the above works.

Generative Representation Learning: Generative models have been applied for learning representations of multimodal data [59, 52]. In particular, [66, 60, 10] proposed a notion of cyclic consistency for learning from unpaired multimodal data using GANs [17], which was extended later to normalizing flows [19]. While these works focus on regularizing a generative mapping between modalities, our notion of cycle consistency applies to embeddings learned via a contrastive framework.

6 Conclusion

We presented CyCLIP, a framework for cycle consistent multimodal representation learning for image and text modality. The main benefits of CyCLIP stem from including cross-modal consistency and in-modal consistency regularizers to prevent inconsistent inference in the image and text spaces. Empirically, we show that CyCLIP performs much better than CLIP on zero-shot classification and is more robust on benchmarks for distributional robustness. We also showed that the representations learned by CyCLIP are more consistent than CLIP and better capture concept-level knowledge, as evidenced by our analysis of fine-grained and coarse-grained accuracies.

We believe this work can motivate further studies on understanding the geometry of the representation spaces learned via the contrastive objective applied to paired multimodal data and, in particular, identify conditions and regularization strategies under which the learned representations are synergistic across the various modalities for downstream applications.

One important future direction and a current limitation is scaling CyCLIP to larger datasets. While we do not possess the resources for this study, it is imperative to study the extent to which the benefits of cycle consistency remain at the scale on which the original CLIP was trained (400M image-text pairs). Finally, for real-world deployment of CLIP and their variants, such as CyCLIP, we need to be cautious about amplifying societal biases as these models are trained on large uncurated datasets scraped from the web [9]. Additionally, it is easy to add malicious data to the web, which poses a severe security threat [5]. Alleviating such harms is an important and active area of research.

References

[1] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. See, hear, and read: Deep aligned representations. *arXiv preprint arXiv:1706.00932*, 2017.

[2] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018.

[3] Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv preprint arXiv:2105.04906*, 2021.

[4] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

[5] Nicholas Carlini and Andreas Terzis. Poisoning and backdooring contrastive learning. *arXiv preprint arXiv:2106.09667*, 2021.

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

[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[8] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

[9] Jaemin Cho, Abhay Zala, and Mohit Bansal. Dall-eval: Probing the reasoning skills and social biases of text-to-image generative transformers. *arXiv preprint arXiv:2202.04053*, 2022.

[10] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8789–8797, 2018.

[11] Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 539–546. IEEE, 2005.

[12] Katherine Crowson. Vqgan-clip: https://colab.research.google.com/drive/1zaus_gn2rhtzwzowuppernc0q8ohzrtz.

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

[14] Karan Desai and Justin Johnson. Virtex: Learning visual representations from textual annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11162–11173, 2021.

[15] Jiali Duan, Liqun Chen, Son Tran, Jinyu Yang, Yi Xu, Belinda Zeng, Chenyang Tao, and Trishul Chilimbi. Multi-modal alignment using representation codebook. *arXiv:2203.00048*, 2022.

[16] Tianyu Gao, Xingcheng Yao, and Danqi Chen. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*, 2021.

[17] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[18] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in Neural Information Processing Systems*, 33:21271–21284, 2020.

[19] Aditya Grover, Christopher Chute, Rui Shu, Zhangjie Cao, and Stefano Ermon. Alignflow: Cycle consistent learning from multiple domains via normalizing flows. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 4028–4035, 2020.

[20] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 297–304. JMLR Workshop and Conference Proceedings, 2010.

[21] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE, 2006.

[22] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

[23] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8340–8349, 2021.

[24] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Lixuan Zhu, Samyak Parajuli, Mike Guo, Dawn Xiaodong Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8320–8329, 2021.

[25] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15262–15271, 2021.

[26] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Xiaodong Song. Natural adversarial examples. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15257–15266, 2021.

[27] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, 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.

[28] 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 *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021.

[29] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015.

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

[31] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[32] 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:2110.05208*, 2021.

[33] Sijie Mai, Ying Zeng, Shuangjia Zheng, and Haifeng Hu. Hybrid contrastive learning of tri-modal representation for multimodal sentiment analysis. *arXiv:2109.01797*, 2021.

[34] 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.

[35] Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. Slip: Self-supervision meets language-image pre-training. *arXiv:2112.12750*, 2021.

[36] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv:2112.10741*, 2021.

[37] Curtis G Northcutt, Anish Athalye, and Jonas Mueller. Pervasive label errors in test sets destabilize machine learning benchmarks. *arXiv preprint arXiv:2103.14749*, 2021.

[38] Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep metric learning via lifted structured feature embedding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4004–4012, 2016.

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

[40] Hieu Pham, Zihang Dai, Golnaz Ghiasi, Hanxiao Liu, Adams Wei Yu, Minh-Thang Luong, Mingxing Tan, and Quoc V Le. Combined scaling for zero-shot transfer learning. *arXiv preprint arXiv:2111.10050*, 2021.

[41] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015.

[42] Ariadna Quattoni, Michael Collins, and Trevor Darrell. Learning visual representations using images with captions. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007.

[43] 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 *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[44] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[45] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

[46] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021.

[47] 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.

[48] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[49] Aditya Sanghi, Hang Chu, Joseph G Lambourne, Ye Wang, Chin-Yi Cheng, and Marco Fumero. Clip-forge: Towards zero-shot text-to-shape generation. *arXiv preprint arXiv:2110.02624*, 2021.

[50] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[51] 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.

[52] Yuge Shi, Brooks Paige, Philip Torr, et al. Variational mixture-of-experts autoencoders for multi-modal deep generative models. *Advances in Neural Information Processing Systems*, 32, 2019.

[53] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. *arXiv preprint arXiv:2112.04482*, 2021.

[54] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. *Advances in neural information processing systems*, 29, 2016.

[55] Ajinkya Tejankar, Bichen Wu, Saining Xie, Madian Khabsa, Hamed Pirsiavash, and Hamed Firooz. A fistful of words: Learning transferable visual models from bag-of-words supervision. *arXiv:2112.13884*, 2021.

[56] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. *Advances in Neural Information Processing Systems*, 32, 2019.

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

[58] Wenhui Wang, Hangbo Bao, Li Dong, and Furu Wei. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *arXiv preprint arXiv:2111.02358*, 2021.

[59] Mike Wu and Noah Goodman. Multimodal generative models for scalable weakly-supervised learning. *Advances in Neural Information Processing Systems*, 31, 2018.

[60] Zili Yi, Hao Zhang, Ping Tan, and Minglun Gong. Dualgan: Unsupervised dual learning for image-to-image translation. In *Proceedings of the IEEE international conference on computer vision*, pages 2849–2857, 2017.

[61] Xin Yuan, Zhe Lin, Jason Kuen, Jianming Zhang, Yilin Wang, Michael Maire, Ajinkya Kale, and Baldo Faieta. Multimodal contrastive training for visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6995–7004, 2021.

[62] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pages 12310–12320. PMLR, 2021.

[63] Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. Merlot: Multimodal neural script knowledge models. *Advances in Neural Information Processing Systems*, 34, 2021.

[64] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5579–5588, 2021.

[65] Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. Contrastive learning of medical visual representations from paired images and text. *arXiv preprint arXiv:2010.00747*, 2020.

[66] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] Section 6.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] Section 6.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No] Due to extremely-compute heavy experiments.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [No] All the non-proprietary datasets and code used are public under MIT, BSD or CC licenses.
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Additional Results

In addition to CyCLIP described in §2, we train two more instantiations of it by keeping either of the two consistency regularizers active in the loss objective (Eq. 5). The instantiation trained by setting $\lambda_1 = 0$ and $\lambda_2 = 0.5$ is termed as C-CyCLIP as only cross-modal consistency regularizer term is added to the loss objective. Similarly, we get I-CyCLIP where only in-modal consistency regularizer is added to the loss by setting $\lambda_1 = 0.5$ and $\lambda_2 = 0$. We evaluate C-CyCLIP and I-CyCLIP on most of the experiments discussed in the main text to understand their zero-shot transfer ability on standard datasets and robustness to natural distribution shifts.

A.1 Zero-shot Transfer

Table 6 presents our results of the zero-shot transfer experiment described in §3.1. We find that CyCLIP outperforms its sub-variants and the CLIP model on the ImageNet1K dataset. Interestingly, we observe that the C-CyCLIP model performs the best amongst all models on CIFAR-10 and CIFAR-100. Further, we notice that all the versions of CyCLIP (the last three rows of Table 6) are better than CLIP across all the datasets. This indicates that jointly improving the geometry of the learned image and text representations using cyclic consistency regularizers is practical for improved zero-shot transfer performance on the image classification task.

Table 6: Zero-shot TopK classification accuracy (%) where K$\in \{1, 3, 5\}$

	CIFAR-10			CIFAR-100			ImageNet1K		
	Top1	Top3	Top5	Top1	Top3	Top5	Top1	Top3	Top5
CLIP	46.54	78.22	91.16	18.69	34.72	43.97	20.03	33.04	39.35
CyCLIP	51.45	79.57	**91.80**	23.15	41.46	50.66	**22.08**	**35.98**	**42.30**
C-CyCLIP	**55.52**	**80.93**	90.60	**25.49**	**44.82**	**54.10**	21.74	35.48	41.96
I-CyCLIP	50.97	79.51	90.75	22.35	39.25	48.45	21.22	34.72	41.05

A.2 Natural Distribution Shifts

We evaluate the performance of the ablated CyCLIP models on natural distribution shift benchmarks to evaluate their distributional robustness and present our results in Table 7. We find that all the CyCLIP models (the last three rows) outperform CLIP by a large margin across all the datasets. Interestingly, we observe that C-CyCLIP performs the best on three of the four natural distribution shift datasets (last three columns). Among the CyCLIP models, I-CyCLIP performs the worst across all the datasets. This indicates that having just in-modal consistency is not enough to preserve the distributional robustness on the traditional datasets.

Table 7: Zero-shot classification on Natural Distribution Shifts (%)

	ImageNetV2			ImageNetSketch			ImageNet-A			ImageNet-R		
	Top1	Top3	Top5	Top1	Top3	Top5	Top1	Top3	Top5	Top1	Top3	Top5
CLIP	16.91	29.28	34.99	10.37	19.15	24.20	4.23	11.35	16.88	24.32	39.69	47.20
CyCLIP	**19.22**	**32.29**	**38.41**	12.26	22.56	28.17	5.35	13.53	19.51	26.79	42.31	50.03
C-CyCLIP	18.65	31.81	38.29	**12.77**	**22.75**	**28.26**	**5.59**	**14.65**	**21.35**	**27.99**	**43.66**	**50.99**
I-CyCLIP	18.40	30.63	36.82	10.87	20.17	25.49	4.75	11.84	16.92	24.55	38.71	45.68

A.3 Linear Probe

We perform linear probing to assess the performance of CLIP and CyCLIP models in the presence of extra in-domain and extra in-modality supervision, i.e., with access to the training datasets for the visual classification task. We first get the image embeddings of the training data points from our trained visual encoder (ResNet-50), followed by training a learnable classifier (linear mapping).

Further, we train our models from scratch for 32 epochs on a single RTX2080Ti GPU with a batch size of 16 and an initial learning rate of 0.005 with cosine scheduling. We use a weight decay of 0.01 for the non-bias parameters of the linear layer.

We experiment with 14 visual classification datasets, details of which are present in Table 8. We present our linear probing results in Table 9. In particular, we observe that the CyCLIP models marginally outperform the CLIP model on all the datasets except Flowers102 and OxfordIIITPet.

Table 8: Training data size, Testing data size, and the number of classes for different datasets used for Linear Probe evaluation.

Dataset	Classes	Train size	Test size
Caltech101	102	3060	6084
CIFAR-10	10	50000	10000
CIFAR-100	100	10000	10000
DTD	47	3760	1880
FGVCAircraft	100	6667	3333
Flowers102	102	2040	6149
Food101	101	75750	25250
GTSRB	43	26640	12630
ImageNet1K	1000	50000	50000
OxfordIIITPet	37	3680	3669
RenderedSST2	2	6920	1821
StanfordCars	196	8144	8041
STL10	10	5000	8000
SVHN	10	73257	26032

Table 9: Transfer CLIP and CyCLIP to 14 downstream visual datasets using linear probing. For training ImageNet1K, we use a random subset of 50K images from its original training dataset. Results have been averaged over 5 different seeds used for training the linear classifier.

	Caltech101	CIFAR-10	CIFAR-100	DTD	FGVCAircraft	Flowers102	Food101	GTSRB	ImageNet1K	OxfordIIITPet	RenderedSST2	StanfordCars	STL10	SVHN	Average
CLIP	79.80	78.26	54.85	59.02	28.00	**83.50**	54.44	69.72	35.93	**57.66**	53.82	20.00	89.23	47.28	57.96
CyCLIP	80.33	76.98	55.74	**63.44**	27.86	82.96	54.96	**71.70**	37.12	56.82	53.74	**22.14**	**90.10**	**48.01**	58.71
C-CyCLIP	**80.83**	**79.15**	**56.15**	61.13	**28.25**	82.14	**56.68**	70.33	**37.81**	57.31	**56.33**	22.05	90.05	46.81	**58.93**
I-CyCLIP	80.73	77.77	55.48	61.87	27.46	81.85	54.00	68.26	36.86	57.44	53.60	20.83	89.95	46.89	58.07

A.4 Consistency in Image and Text Spaces

Table 10 presents the results for the consistency score defined in Equation 2. We find that C-CyCLIP is the most consistent model despite missing the in-modal consistency regularizer term. Additionally, we observe that all the CyCLIP models (last three rows) are more consistent than the CLIP model across all the datasets. Hence, the explicit geometrization of the representations helps in improving the consistency of the image and text spaces.

A.5 Fine-grained and Coarse-grained Performance

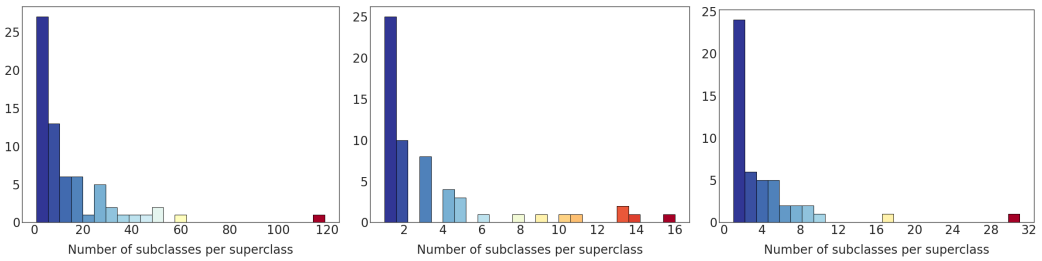
In §4.2, we described a novel fine-grained and coarse-grained analysis to investigate the zero-shot transfer phenomena better. We have 1000, 200, and 200 subclasses distributed across 67, 59, and 49 superclasses for the ImageNet1K/V2/Sketch, ImageNet-A, and ImageNet-R datasets, respectively. Figure 5 illustrates the distribution of the subclasses across the superclasses for these ImageNet

Table 10: Consistency score (%) trend for CLIP and CyCLIP across standard benchmarks . Top-k consistency score implies the fraction of times, the zero-shot predicted label in the text space is identical to the k-Nearest Neighbor predicted label in the image space (using the training dataset).

	CIFAR-10				CIFAR-100				ImageNet1K			
	Top1	Top3	Top5	Top10	Top1	Top3	Top5	Top10	Top1	Top3	Top5	Top10
CLIP	44.60	46.04	47.06	48.45	16.21	17.28	18.42	19.36	16.34	17.42	18.58	19.78
CyCLIP	48.81	50.89	52.30	53.71	20.43	21.96	23.18	24.31	19.20	20.31	21.95	23.94
C-CyCLIP	**53.47**	**56.56**	**57.84**	**59.27**	**22.72**	**24.03**	**25.85**	**26.95**	**20.02**	**21.30**	**23.06**	**24.75**
I-CyCLIP	48.34	49.98	51.27	52.55	20.32	21.32	22.56	23.65	18.82	19.81	21.33	22.92

benchmarks. For the CIFAR-100 dataset, we have 100 subclasses uniformly distributed across 20 superclasses, i.e., 5 subclasses per superclass[1].

Figure 6 illustrates the performance of CLIP and CyCLIP models across the traditional datasets and their natural distribution shifted variants. In tandem with our findings in §4.2, we observe that even ablated CyCLIP models, I-CyCLIP and C-CyCLIP outperform the CLIP model on the coarse-grained and fine-grained classification metric. Additionally, the margin of improvement is larger in the case of coarse-grained analysis than the fine-grained one. This highlights that all the CyCLIP variants are better at capturing entity-level knowledge than CLIP in their joint representations. We also find that CyCLIP captures more entity-level knowledge than C-CyCLIP in ImageNet1K and ImageNetV2, datasets containing natural images belonging to large number of categories.



(a) ImageNet1K/V2/Sketch (b) ImageNet-A (c) ImageNet-R

Figure 5: Distribution of superclasses across different number of subclasses. ImageNet1K/V2/Sketch and ImageNet-A/R have 1000 and 200 subclasses respectively.

A.6 Zero-shot Transfer: Additional Datasets

We present the results for zero-shot transfer downstream task, described in §3.1 in Table 11 across additional image classification datasets. The list of additional datasets is almost identical to the one used for linear probing §9. We include an additional dataset, *SUN397*, with 108754 images split across 397 classes. Further, we do not include some visual datasets as their labels were not amenable to constructing a simple proxy caption. For instance, *GTSRB* is a visual dataset with German traffic signs where a label is decided by the associated color, shape, and sign ID. Likewise, the images in the FGVCAircraft dataset are described by multiple attributes, and those in the RenderedSST2 dataset are described by positive/negative sentiments, hence not relevant in this setting. Furthermore, we present our results on a cleaner test set of CIFAR-10, CIFAR100, and ImageNet1K [37]. We find that CyCLIP outperforms CLIP on the Top-1 classification accuracy with an average gain of 10.6% across all the datasets.

[1] https://www.cs.toronto.edu/ kriz/cifar.html

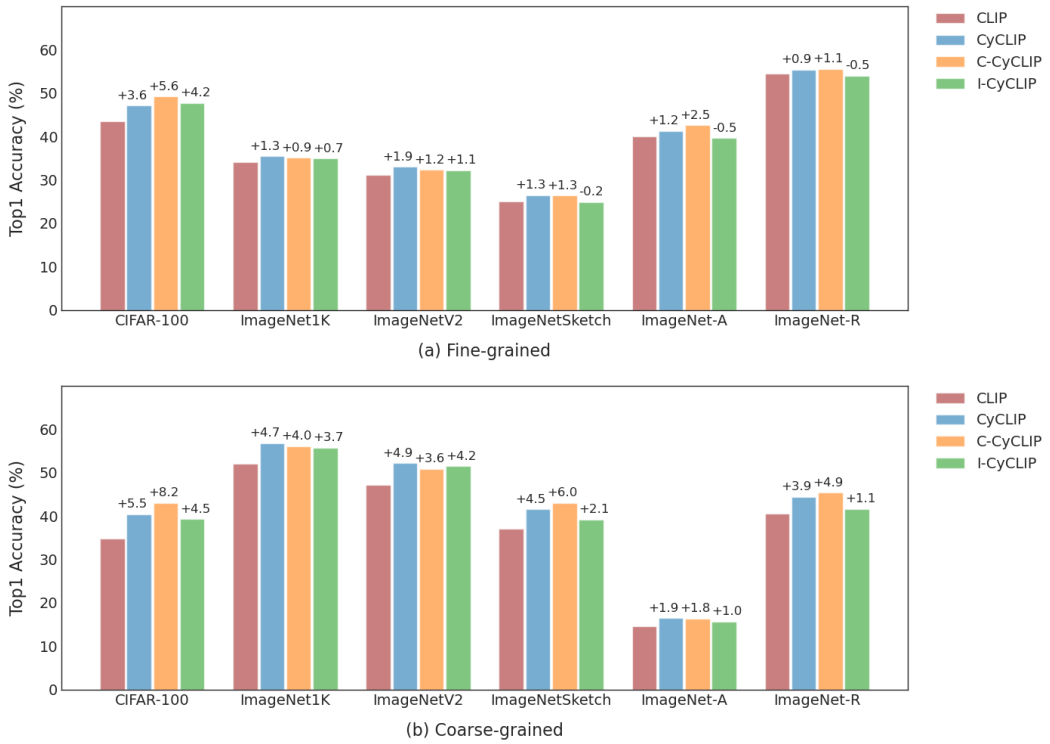


Figure 6: The gap between the performances of CLIP and our CyCLIP variants is much larger in coarse-grained scenario highlighting better entity-level knowledge representation in CyCLIPs.

Table 11: Zero-shot Top1 classification accuracy (%) across a battery of visual datasets. For the datasets marked with (*) we use their error-free labels instead of the original labels in their test sets.

	Caltech101	CIFAR-10*	CIFAR-100*	DTD	Flowers102	Food101	ImageNet1K*	OxfordIIITPets	StanfordCars	STL10	SVHN	SUN397	Average
CLIP	52.6	46.3	18.4	14.7	**13.7**	13.3	20.2	2.2	1.2	83.5	7.1	39.6	26.1
CyCLIP	**60.7**	**50.8**	**22.0**	**19.2**	10.9	**14.0**	**22.5**	**2.3**	**1.6**	**86.1**	**13.3**	**42.7**	**28.8**
Gain (%)	15.4	9.8	19.1	30.0	-20.4	5.0	11.8	7.3	34.4	3.1	86.3	7.7	10.6

A.7 Image-Text Retrieval

We evaluate the effectiveness of the proposed method on the cross-modal (image to text and text to image) retrieval downstream task in the zero-shot as well as fine-tuned settings. We consider the standard benchmark datasets: Flickr30K [41] and MSCOCO [8]. We assess our models on the test set of Flickr30K (1K) and MSCOCO (5K) obtained from the well-known Karpathy [29] split. Both the datasets contains 5 paired captions per image that makes text retrieval per image more easier than image retrieval per caption. We confirm the same in our results below.

We perform fine-tuning on the Karpathy's training split with the batch size of 48. We fine-tune on Flick30K for 10 epochs and MSCOCO for 5 epochs. All the other hyperparameters are identical to that of pre-training.

Table 12 presents our cross-modal image-text retrieval results for CLIP and CyCLIP. In the zero-shot setting, we find that CyCLIP marginally outperforms CLIP on the image retrieval task on both the

Table 12: Zero-shot and fine-tuned cross-modal image-text retrieval (text-to-image and image-to-text) results of CLIP and CyCLIP on Flick30K and MSCOCO datasets.

		Flickr30K (1K)						MSCOCO (5K)					
		Text Retrieval			Image Retrieval			Text Retrieval			Image Retrieval		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
Zero-shot	CLIP	88.2	93.9	95.8	29.9	57.2	68.0	82.1	85.6	87.8	8.4	19.5	26.6
	CyCLIP	88.1	93.7	95.9	**30.9**	57.8	**69.1**	82.1	85.6	87.7	8.6	20.0	27.0
Fine-tuned	CLIP	91.9	97.0	98.0	46.3	74.7	83.6	83.2	87.6	90.0	10.6	23.9	31.3
	CyCLIP	92.3	97.0	98.4	**47.3**	**76.6**	**85.4**	83.2	87.8	90.3	**11.4**	**25.8**	**33.4**

datasets. The relatively lower performance of both CLIP and CyCLIP in the zero-shot setting may be attributed to the more complicated nature of the two datasets where the models are expected to find similarities between the image and text at multiple resolutions as opposed to image classification where there is mostly single object to be matched with a simpler caption. It is not clear as to what distinctions in the raw input and text space are reflected in the embedding space too. Hence, we perform fine-tuning on both the datasets to better inform our models of the downstream datasets. In the fine-tuning setting, we find that the performance of both the models increases across both the datasets. However, we observe clear benefits of the soft consistent regularization on the image retrieval results for both the datasets.

B Additional Discussion

For the models pre-trained on the web-crawled data, one might argue against using the in-modal and cross-modal regularization as the image-text paired data is noisy where the caption may not fully describe the image at all resolutions. However, in an unsupervised setting, it is not pronounced as to what extent these distinctions should be reflected in the embedded representations.

As an example, we might think of the potential harm in bringing the embeddings of the two seemingly different captions ('A person playing with a dog on a beach', 'A person playing with a cat in a room') describing similar-looking images (with 'dog, person, beach, ball, sky', with 'cat, person, room') close using the in-modal consistency constraint. However, if the downstream task is to classify cats vs dogs, then the presence of any other objects in the images is a spurious correlation to be ignored. In such cases, we would desire the in-modal distance between text and image embeddings to be similar for consistent predictions in both modalities. Hence, the effectiveness of any information captured in a modality depends on the downstream task.

Moreover, one could make a similar counter-argument about the text and image modalities being inherently different against CLIP's contrastive loss. It weighs each image and text pair equally, even though some text captions might be more descriptive about the images than others and therefore should be assigned a higher weight. From a practical standpoint, our evidence suggests that soft consistency regularization in the form of additional loss terms, as in Equation 5, can be generally helpful for the downstream tasks and domain settings of interest. The relative gains can indeed be different across tasks and domains. For example, in Figure 3, while our regularizers lead to gains in fine and coarse-grained classification, we noted that the relative gains are much higher for coarse-grained settings due to the improved consistency.

C Pretraining and Implementation details

C.1 Dataset

Conceptual Captions 3M (CC3M) [51] is an open-source dataset consisting of approximately 3.3M image-caption pairs scarped from the web. The dataset (including train/validation split) is made

available by Google [2]. Due to the broken image URLs, we use a subset of 2,631,703 image-caption pairs for pre-training. Further, Conceptual 12M (CC12M) [6] is a noisy extension to the CC3M dataset and contains approximately 12M image-caption pairs, covering a more diverse set of visual concepts. For the dataset ablation experiments in Figure 4 (a), with a data size of 3M pairs, we extend our CC3M dataset using a random subset of image-caption pairs (368,297 pairs) from CC12M [3].

C.2 Model architecture

Our models use the same architecture as the original CLIP model presented in [43] with a ResNet-50 image encoder (38,316,896 parameters) and a transformer-based text encoder with a projection layer (63,690,240 parameters) to match the image embedding dimension of 1024. We use a weight decay for all the parameters during training, except for batch/layer norm, bias, and logit scale parameters.

C.3 Hyperparameter settings

The hyperparameters for the best-performing configuration were inspired from the original paper on CLIP [43] and mainly taken from mlfoundations[4] repository. However, in our experiments, we use a batch size of 128 due to limited computation resources. To avoid hurting the performance of our models, we train our models for 64 epochs. For the cyclic consistency hyperparameters (λ_1, λ_2), we used a combination of zoom grid search and manual tuning on a small subset of 480K image-text pairs with training on 16 epochs and optimizing the contrastive loss on the Conceptual Captions validation set of 13,156 examples. We found that CyCLIP loss is not very sensitive in the non-zero parameter space between 0 and 1. Therefore, we choose a more simplistic setup with both the values as 0.25. We trained the CLIP and CyCLIP models on an Azure cluster with 4x Tesla-V100-SXM2-16GB GPUs and 24x CPUs for approximately 84 hours each. The hyperparameter settings for all our models are shown in table 13.

Table 13: Hyperparameters used for training the CLIP models

Hyperparameter	Value
Embedding dimension	1024
Logit scale range	0 to 4.6052
Epochs	64
Batch size	128
Learning rate	0.0005
Optimizer	Adam
Adam beta1	0.9
Adam beta2	0.99
Adam epsilon	1e-8
Adam weight decay	0.1
Scheduler	Cosine
Learning rate warmup steps	10000

C.4 Code and Reproducibility

The code is made available here. The instructions to download the data and train the models are provided in the ReadMe. For evaluation purposes, the trained checkpoints for CLIP, CyCLIP, C-CyCLIP and I-CyCLIP have also been provided.

[2] https://ai.google.com/research/ConceptualCaptions/download/
[3] https://github.com/google-research-datasets/conceptual-12m
[4] https://github.com/mlfoundations/open_clip