Non-Contrastive Learning Meets Language-Image Pre-Training

Jinghao Zhou Li Dong Zhe Gan Lijuan Wang Furu Wei
Microsoft

Abstract

Contrastive language-image pre-training (CLIP) serves as a de-facto standard to align images and texts. Nonetheless, the loose correlation between images and texts of web-crawled data renders the contrastive objective data inefficient and craving for a large training batch size. In this work, we explore the validity of non-contrastive language-image pre-training (nCLIP), and study whether nice properties exhibited in visual self-supervised models can emerge. We empirically observe that the non-contrastive objective benefits representation learning while sufficiently underperforming under zero-shot recognition. Based on the above study, we further introduce xCLIP, a multi-tasking framework combining CLIP and nCLIP, and show that nCLIP aids CLIP in enhancing feature semantics. The synergy between two objectives lets xCLIP enjoy the best of both worlds: superior performance in both zero-shot transfer and representation learning. Systematic evaluation is conducted spanning a wide variety of downstream tasks including zero-shot classification, out-of-domain classification, retrieval, visual representation learning, and textual representation learning, showcasing a consistent performance gain and validating the effectiveness of xCLIP. The code and pre-trained models will be publicly available at https://github.com/shallowtoil/xclip.

1. Introduction

Language-image pre-training which simultaneously learns textual and visual representation from large-scale image-text pairs has revolutionized the field of representation learning [25, 59], vision-language understanding [22], and text-to-image generation [61]. Compared to traditional visual models, language-instilled ones intrinsically inherit the capability of zero-shot or few-shot learning prominently demonstrated by large language models such as GPT-3 [12]. The precursor system, **C**ontrastive **L**anguage-**I**mage **Pre**-Training [59] (**CLIP**) that explicitly aligns the projected features of two modalities, has demonstrated surprising capabilities of zero-shot, representation learning, and robustness, being applied to a wide range of fields [32, 52, 60, 70].

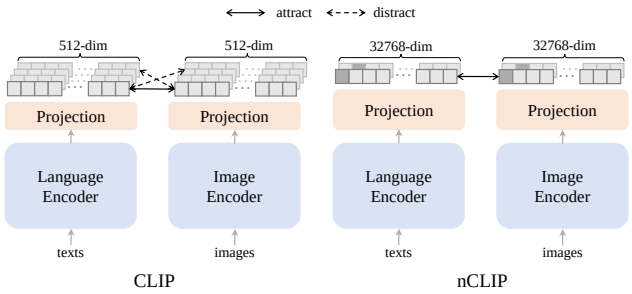

Figure 1. **Architecture comparison between CLIP and nCLIP.** We take **base**-size encoders for an instance. CLIP discriminates instances within a batch using 512-dim projected embeddings. nCLIP projects each modality into a 32768-dim probability distribution as the pseudo-label to supervise the prediction from the other modality. Darker blocks of nCLIP depict a higher response for cluster distribution, signifying the clusters to which text or image instances may belong. xCLIP is a multi-tasking framework with both CLIP and nCLIP.

To learn from noisy web-crawled image-text pairs, CLIP adopts a formulation of the contrastive objective, where the image and text within a pair are considered as unique instances and are encouraged to be discriminated from all the other negative instances. However, web-crawled image-text pairs [65, 69] are usually loosely correlated in the sense that one caption (image) can match reasonably with multiple images (captions) besides the ground-truth one, as statistically shown in [74]. Hence, it is inaccurate and data-inefficient for representation learning to neglect other sensible matches and overlook the semantics hidden inside the textual description. This is also solidified by the undesirable discriminative ability or transferring performance of the visual encoder pre-trained under the contrastive objective, as suggested in [78]. Mining semantics from plentiful concepts appearing in captions, therefore, solicits further exploration beyond the vanilla contrastive objective.

Previously, some works resort to mining nearest neighbor positive samples via intra-modality similarity [50, 74] but require extra storage for auxiliary modules, *i.e.*, the teacher network [74] or the memory bank [50]. Other works conduct multi-tasking with image-label supervision [57, 78, 80], transforming the captions into tag format or the image tag into the captioning format for unified contrastive

learning. Notwithstanding, such practice is still unable to account for the loose assignment within the image-text captioning pairs. Intuitively, a caption for an image usually identifies existing objects within the image, which can be captured by *a probabilistic distribution estimating how the text is assigned to one or multiple object clusters* [13, 14]. Such estimation depicts the semantic meanings for its contained visual contents, and thus can be leveraged as a pseudo-label to guide the learning of the visual encoder.

Inspired by the superiority of visual self-supervised learning (SSL) models pre-trained with the non-contrastive objective [14, 15], we explore whether the non-contrastive objective can be used across modalities for pre-training language-image models, and whether nice properties displayed on visual SSL models can be inherited. To this end, we follow the same setup as CLIP except for the objective, and study **n**on-**C**ontrastive **L**anguage-**I**mage **P**re-Training (**nCLIP**). For an image-text pair, we use the estimation of the textual (visual) distribution as the target to supervise the visual (textual) distribution, measured by cross-entropy. Additional regularizers are applied to avoid trivial solutions. Schematic comparison between CLIP and nCLIP is shown in Fig. 1. Theoretically, such formulation takes one modality and the cluster centroid that the other modality belongs to as the positive samples to contrast [14], as opposed to direct features of two modalities as in contrastive learning, such that a single image is tolerated to be aligned with multiple captions and concepts. Based on the above formulation, we conduct a systematic study in terms of zero-shot transfer and representation learning. We empirically observe that, while nCLIP demonstrates desirable representation learning capability for both modalities, it underperforms prominently in zero-shot classification and retrieval tasks where models pre-trained with negative samples naturally prevail.

Seeking for unique strengths of both worlds, we further perform multi-tasking of CLIP and nCLIP, short as **xCLIP**, and seek synergy between two distinct pre-training objectives. With the bearable add-on of computational resources (*e.g.*, $\sim 27\%$ memory-wise and $\sim 30\%$ time-wise), xCLIP achieves consistent performance gain compared to CLIP on a wide range of tasks spanning from zero-shot classification, retrieval, linear probing, and fine-tuning, *etc*. An extensive study with different pre-training datasets, evaluation metrics, and optimization configurations is conducted to validate xCLIP's effectiveness in mining semantics and improving data efficiency. Particularly, the base-size model pre-trained using xCLIP under 35-million publicly available image-text pairs achieves a performance gain of **3.3**% and **1.5**% on an average of 27 classification tasks [59], in terms of zero-shot classification and linear probing accuracy, respectively. Performance gain is also valid in terms of zero-shot retrieval, semi-supervised learning, and fine-tuning, with **3.7** points of R@1 on Flickr30K [58], **1.7**% accuracy on 1% subset of ImageNet [24], and **0.3**% accuracy on ImageNet, respectively.

2. Related Work

Language-image pre-training. Language-image pre-training, *aka* vision-language pre-training (VLP), learns to jointly cope with vision and language input using multi-modality models. Early practices [22, 49, 67, 68] separately encode modalities, *i.e.*, texts with linear embedding and images with convolutional neural networks and region proposals [63]. Recent works take direct images instead of pre-fetched region features as input, using *e.g.*, dual-encoder architecture for alignment [42, 57, 59], single-encoder architecture for fusion [43, 73], or their combinatorial practices forming an encoder-decoder architecture [48, 80]. Most recent literature aims to expand as many supported transferable downstream tasks as possible via pre-training large-scale foundation models [6, 80]. Out of previous works, CLIP [59] blazes the trail to learn transferable visual features from text supervision, and demonstrates surprising zero-shot recognition results. Follow-up study includes leveraging dense [79], augmentation [47], and uni-modal self-supervised [50, 55, 74] signals for improved performance. In this work, we base our study on the dual-encoder architecture, and pre-train models either with the *contrastive* objective as CLIP [59], or the *non-contrastive* objective. Further, we seek their complementarity and study whether synergy between two objectives exists.

Contrastive learning. The idea of contrastive learning is popularized under visual self-supervised learning (SSL) which aims at learning transferable visual representation from unlabelled images. In common practices [18, 33, 75], each image is considered a single class, attracted to its augmented views while pulled away from views of other images. The contrastive objective is proved robust and effective in a wide range of realms beyond visual understanding, including representation learning for natural language [29], audio [54], structured input [44], robotics [66], and multi-modality [21, 59]. However, a caveat of these approaches is the requirement for large batch size [18] or memory bank [33, 75], which lies intrinsic in the formulation of InfoNCE estimator [38] and plagues the pre-training of large-size models due to hardware limitations.

Non-contrastive learning. Beyond the contrastive paradigm, recent state of the arts from visual SSL manage to relieve the dependency on negative samples. With different subtleties to avert collapsing solutions, these works can optimize the affinity of augmented representations alone

and are categorized as *non-contrastive* framework. To avoid model collapsing, common practices include asymmetrical architecture [19, 31], dimension de-correlation [11, 27, 81], and clustering [7, 8, 13–15, 71]. Our approach takes inspiration from the last category since it intrinsically frees the absoluteness of assignment of positive samples. Moreover, rich semantics induced by explicit clustering can automatically group visual concepts from noisy image-text pairs, facilitating models' representation learning.

3. Approach

3.1. Framework

Suppose $\boldsymbol{g}$ and $\boldsymbol{h} \in \mathbb{R}^{D\times 1}$ are projected backbone features of two modalities. D is the feature dimension. We start by formulating the objective of dominant *contrastive* framework CLIP, followed by the introduced *non-contrastive* framework nCLIP. The final framework xCLIP takes the multi-tasking of two objectives, the complete computational pipeline of which is shown in Algorithm 1.

Contrastive pre-training [59]. Let $\boldsymbol{u} = \boldsymbol{g}/||\boldsymbol{g}||$ and $\boldsymbol{v} = \boldsymbol{h}/||\boldsymbol{h}||$. $\boldsymbol{U} = [\boldsymbol{u}_1, \boldsymbol{u}_2, ..., \boldsymbol{u}_B] \in \mathbb{R}^{D\times B}$, $\boldsymbol{V} = [\boldsymbol{v}_1, \boldsymbol{v}_2, ..., \boldsymbol{v}_B] \in \mathbb{R}^{D\times B}$ are concatenated embeddings over the batch. B is the batch size. The contrastive objective is formulated as

$$\begin{aligned}\mathcal{L}_{\text{CLIP}} &= \text{InfoNCE}(\boldsymbol{U}^{\text{T}}\boldsymbol{V}/\sigma) \\ &= -\frac{1}{B}\sum_{i\in B}\log\frac{\exp(\boldsymbol{u}_i^{\text{T}}\boldsymbol{v}_i/\sigma)}{\sum_{j\in B}\exp(\boldsymbol{u}_i^{\text{T}}\boldsymbol{v}_j/\sigma)},\end{aligned} \tag{1}$$

where σ is a trainable parameter controlling the temperature. $\mathcal{L}_{\text{CLIP}}$ is symmetrized by setting $\boldsymbol{g}$ and $\boldsymbol{h}$ as projected features of the images and texts by turns and taking the average of two terms. Vanilla InfoNCE can be decoupled into two terms accounting for affinity and variability separately [81]. While both variability terms in the contrastive and non-contrastive objectives require for batch statistics (*e.g.*, $\mathcal{L}_{\text{EH}}$ in Eq. (4)), CLIP formulation explicitly maximizes the distance between negative pairs of batch samples. Hence, it degrades the learning performance and data efficiency when models are pre-trained with noisy data where sensible matches occur within negative pairs.

Non-contrastive pre-training. Let $\boldsymbol{p} = \text{softmax}(\boldsymbol{g})$ and $\boldsymbol{q} = \text{softmax}(\boldsymbol{h})$. We transform the projected feature into probability distributions, which can be seen as an assignment over semantic clusters (*i.e.*, projection weights). We take $\boldsymbol{p}$ as the target distribution and learn to estimate the predicted distribution $\boldsymbol{q}$ by minimizing their cross-entropy:

$$\mathcal{L}_{\text{CE}} = -\boldsymbol{p}^{\text{T}}\log(\boldsymbol{q}). \tag{2}$$

Note the target and prediction branch are backpropagated at the same time and $\mathcal{L}_{\text{CE}}$ is symmetrized as $\mathcal{L}_{\text{CLIP}}$. To avert collapsing solutions, we incorporate entropy regularizers, *i.e.*, entropy minimization [71] and mean entropy maximization [8, 9], via minimizing:

$$\mathcal{L}_{\text{EH}} = -\boldsymbol{p}^{\text{T}}\log(\boldsymbol{p}),\ -\mathcal{L}_{\text{HE}} = \overline{\boldsymbol{p}}^{\text{T}}\log(\overline{\boldsymbol{p}}), \tag{3}$$

where $\overline{\boldsymbol{p}} = \mathbb{E}(p) \approx \frac{1}{B}\sum_{i=1}^{B}\boldsymbol{p}_i$ is the average distribution over batch. $\mathcal{L}_{\text{EH}}$ and $\mathcal{L}_{\text{HE}}$ are symmetrized by taking the average with another term where $\boldsymbol{p}$ is replaced by $\boldsymbol{q}$. $\mathcal{L}_{\text{EH}}$ encourages the model to make deterministic predictions and ensures assignment's *sharpness*. $\mathcal{L}_{\text{HE}}$ encourages the model to utilize a full set of projection weights and ensures the assignment's *smoothness*. We illustrate in Sec. 4.4 and Appendix A that $\mathcal{L}_{\text{EH}}$ and $\mathcal{L}_{\text{HE}}$ are minimally sufficient to yield non-collapsing solutions. We empirically show in Tab. A1 that the solutions is non-trivial as long as both *sharpness* and *smoothness* are guaranteed, *e.g.*, via Sinkhorn algorithm [14] instead of loss regularizers. We opt for Eq. (3) for its simplicity and consistency between pre-training and evaluation. The overall non-contrastive objective is formulated as

$$\mathcal{L}_{\text{nCLIP}} = \mathcal{L}_{\text{CE}} + \lambda_1\cdot\mathcal{L}_{\text{EH}} - \lambda_2\cdot\mathcal{L}_{\text{HE}}, \tag{4}$$

where λ_1 and λ_2 controls the weight of the regularization. We empirically find that setting $\lambda_1 = \lambda_2 - 1 = 0.5$ yields stable training and favorable transfer performances. We detail our findings in Sec. 4.4 and Appendix B.1. During evaluation, we utilize negative cross-entropy $-\mathcal{L}_{\text{CE}}$ as the similarity metric for zero-shot transfer experiments. As schematically showcased in Fig. 1, nCLIP does not reply on negative examples but requires a relatively larger projection head and greater output dimension. Comparing nCLIP with CLIP, we empirically find that nCLIP produces more coarse-grained (*e.g.*, zero-shot retrieval in Sec. 4.1) but semantic-rich (*e.g.*, linear probing in Sec. 4.2) projections. Empirical evidence suggests that the model will fail to deliver reasonable zero-shot and retrieval results if pre-trained without negative pairs (*e.g.*, nCLIP's 25.2 *vs.* CLIP's 73.8 points of R@1 under Flickr30K I→T retrieval as shown in Tab. 4).

Contrastive meets non-contrastive. Given that two objectives each have their own limitations, we further seek the complementarity between $\mathcal{L}_{\text{CLIP}}$ and $\mathcal{L}_{\text{nCLIP}}$, and pre-train the models with both objectives, written as

$$\mathcal{L}_{\text{xCLIP}} = \lambda_{\text{CLIP}}\cdot\mathcal{L}_{\text{CLIP}} + \lambda_{\text{nCLIP}}\cdot\mathcal{L}_{\text{nCLIP}}, \tag{5}$$

where λ_{CLIP} and λ_{nCLIP} control the weight of the objectives. During evaluation, we find using the negative cosine metric with CLIP's projection head for zero-shot transfer

Algorithm 1: PyTorch-like pseudo-code of xCLIP.

```
// enc, proj:  encoder & projector
f_I, f_T = enc(img), enc(text)
g_I, g_T = proj_CLIP(f_I), proj_CLIP(f_T)
h_I, h_T = proj_nCLIP(f_I), proj_nCLIP(f_T)
L_xCLIP = λ_CLIP · L_CLIP(g_I, g_T, σ)
        +λ_nCLIP · L_nCLIP(h_I, h_T, λ_1, λ_2)
return L_xCLIP

def L_CLIP (f_I, f_T, σ):
    // INCE: InfoNCE with Eq. (1)
    u_I = normalize(f_I, p=2, dim=1)
    u_T = normalize(f_T, p=2, dim=1)
    L = INCE(U_I^T U_T/σ) + INCE(U_T^T U_I/σ)
    return L / 2

def L_nCLIP (f_I, f_T, λ_1, λ_2):
    p_I = softmax(f_I, dim=1)
    p_T = softmax(f_T, dim=1)
    L_CE = - (p_I · log(p_T) + p_T · log(p_I)).sum(dim=1)
           .mean(dim=0)
    L_EH = - (p_I · log(p_I) + p_T · log(p_T)).sum(dim=1)
           .mean(dim=0)
    p̄_I, p̄_T = p_I.mean(dim=0), p_T.mean(dim=0)
    L_HE = - (p̄_I · log(p̄_I) + p̄_T · log(p̄_T)).sum(dim=1)
    L = L_CE + λ_1 · L_EH - λ_2 · L_HE
    return L / 2
```

experiments generally leads to better results. We note that it is not immediately clear that the two objectives certainly induce stronger models, since they build qualitatively distinctive latent spaces. Qualitatively, we show in Sec. 4 that nCLIP helps CLIP to encode semantics which intrinsically lacks in CLIP, while CLIP helps nCLIP to be transferable for zero-shot recognition. The synergy between the two objectives prompts consistent performance gains across a wide range of tasks by xCLIP over CLIP.

3.2. Implementation

Architecture. We train the base-size model with ViT-B/16 [26] as the visual encoder. The model configuration of the base-size text encoder follows CLIP [59] with byte-pair encoding (BPE) and a maximum context length of 77. The projection head generating output for the non-contrastive objective is a 2-layer MLP with 4096-dim hidden layers, GELU [36], and BatchNorm [41]. The last layer is of 32768 output dimension and followed by a BatchNorm without affine transformation. The projection head for the contrastive objective is a single linear layer with no bias and a dimension of 512.

Pre-training data. We train our method with publicly available datasets: COCO [51], Visual Genome [45], SBU Captions [56], Conceptual Caption 3M [65], Conceptual Caption 12M [17], and filtered 14M-size subset of Yahoo Flickr Creative Commons 100M dataset, consisting a total of 35M Image-Text pairs (IT35M). We are also intrigued about the behaviors of models pre-trained on ImagNet-21K [24] (IN21K) dataset taking label names as annotation texts by concatenating them with a sampled prompt, similar to [57, 80]. To notify, ImageNet-21K is a subset of ImageNet-22K with classes of ImageNet-1K excluded for fair downstream evaluation. We detail the models' data scaling behavior in Tab. 8.

Optimization. We use a batch size of 4096, with all data distributed across 32 V100 GPUs. Models are trained by default with AdamW [53] optimizer, a peak learning rate of $1e^{-3}$, cosine scheduler, a weight decay of 0.2, a β_2 for AdamW of 0.98, a ϵ for AdamW of $1e^{-6}$, and automatic mixed-precision for 3 warm-up epochs and a total of 32 epochs. λ_1 and λ_2 are set as 0.5 and 1.5, respectively. The two objectives are multi-tasked with λ_{nCLIP} being 1 and λ_{CLIP} being 0.2. Augmentations and pre-processing of pre-training data follow CLIP with image size randomly cropped and resized within the scale of (0.5, 1.0). Detailed pre-training hyper-parameters are detailed in Appendix F.

4. Experiments

4.1. Zero-Shot Transfer

Classification. We evaluate 27 classification benchmarks with zero-shot classification protocols following [59]. As shown in Tab. 1, nCLIP achieves an average top-1 accuracy of 32.7%. Though performing 7.6% worse than CLIP, nCLIP demonstrates descent zero-shot recognition capability without explicitly training with negative examples. xCLIP achieves consistent performance gain across a wide range of datasets, with a 43.6% average top-1 accuracy, which is **3.3**% higher than CLIP. It indicates that the synergy between contrastive and non-contrastive objectives improves the model's zero-shot classification ability.

Out-of-domain classification. For out-of-distribution classification, we use 5 datasets following [59]: ImageNet Adversarial [37], ImageNet Rendition [35], ImageNetV2 [62], ImageNet Sketch [72], and ObjectNet [10] with 2 additional datasets: ImageNet-C [76] and Stylized ImageNet [30]. We investigate if the non-contrastive objective inherits the robustness of the contrastive objective to natural distribution shift. As shown in Tab. 2, nCLIP achieves an 18.0% with CLIP an 23.1% average top-1 accuracy, revealing a similar trend to in-domain datasets. xCLIP further improves CLIP's accuracy by **3.6**% and achieves an accuracy of 26.7% across 7 datasets, indicating its effectiveness among out-of-domain datasets.

Model	Food101	CIFAR10	CIFAR100	Birdsnap	SUN397	Cars	Aircraft	VOC2007	DTD	Pets	Caltech101	Flowers	MNIST	FER2013	STL10	EuroSAT	RESISC45	GTSRB	KITTI	Country211	PCAM	UCF101	Kinetics700	CLEVR	HatefulMemes	SST	**ImageNet**	**Average**
CLIP	61.2	79.7	50.6	23.7	56.5	15.9	5.8	46.4	27.6	54.7	71.3	48.9	10.5	37.3	91.3	24.5	39.7	13.1	31.6	9.1	50.0	45.0	32.4	12.8	**53.0**	49.1	45.7	40.3
nCLIP	28.4	79.5	49.1	11.3	57.0	5.9	4.5	51.4	22.9	14.6	65.0	23.1	9.9	13.5	**94.8**	15.1	21.2	2.7	35.4	5.8	**51.2**	42.0	28.4	12.4	52.7	50.0	37.0	32.7
xCLIP	**65.8**	**83.4**	**54.5**	**25.1**	**59.9**	**18.0**	**5.8**	**52.2**	**33.2**	**57.1**	**73.9**	**50.0**	**12.3**	**39.0**	92.8	**40.0**	**43.6**	**16.3**	**39.8**	**9.3**	51.1	**49.8**	**35.4**	**18.4**	52.5	**50.2**	**48.8**	**43.6**

Table 1. **Zero-shot classification.** We report on a variety of classification benchmarks with ViT-B/16 pre-trained on IT35M. Detailed protocols for each dataset strictly follow CLIP [59]. xCLIP achieves a consistent performance gain compared to CLIP in a wide range of classification datasets. Best results of each column are **bolded**.

Model	IN-A	IN-R	IN-v2	IN-Ske	IN-Sty	IN-C	ObjNet	**Average**
CLIP	21.3	43.7	37.9	26.5	4.9	11.9	15.5	23.1
nCLIP	21.5	29.0	30.8	17.7	3.6	11.5	12.0	18.0
xCLIP	**27.7**	**49.1**	**40.6**	**30.6**	**6.1**	**14.7**	**17.8**	**26.7**

Table 2. **Zero-shot out-of-distribution classification.** We report on a variety of out-of-distribution classification benchmarks. xCLIP demonstrates stronger robustness on various out-of-domain classification datasets.

Model	NUS-WIDE			OpenImages		
	mAP	F1@3	F1@5	mAP	F1@10	F1@20
CLIP	15.1	33.3	16.0	81.1	13.4	7.2
nCLIP	**16.1**	**35.6**	16.6	**81.4**	11.6	6.4
xCLIP	15.3	35.2	**16.8**	81.2	**13.8**	**7.4**

Table 3. **Zero-shot multi-label classification.** We report mAP and F1 scores with ViT-B/16 pre-trained on IT35M. nCLIP showcases the best multi-label classification capability.

Multi-label classification. We evaluate zero-shot multi-label classification on NUS-WIDE [23] and OpenImages [46] following standard protocol [39]. Specifically, we use 81 unseen labels for NUS-WIDE and the most frequent 400 unseen test labels for OpenImages during evaluation, following [40]. As shown in Tab. 3, we observe that nCLIP shows the best mAP compared to CLIP. nCLIP is pre-trained without negative examples, resulting in the model's intrinsic strengths on recall over overlapped concepts, with 16.1 and 81.4 points of mAP on NUS-WIDE and OpenImages, respectively. Comparatively, xCLIP slightly improves CLIP in this respect while lagging behind nCLIP. The experiments demonstrate that the non-contrastive objective fits well for those downstream tasks in demand of high recall.

Retrieval. We evaluate on 2 retrieval benchmarks: Flickr30K [58] and MSCOCO [51] under zero-shot protocol. We do not use prompt engineering and use the original caption. Compared to zero-shot classification, image-text retrieval requires the model's recognition capability at a fine-grained level. We empirically observe in Tab. 4 that nCLIP performs drastically worse than CLIP, which is because text captions in retrieval are not mutually exclusive, which is different from label names in classification. Therefore, explicit negative examples during pre-training play an imperative role in zero-shot retrieval. See Appendix E.1 for further discussions. Beyond that, xCLIP achieves a noticeable performance improvement over CLIP, with a gain of **3.7**% R@1 on Flickr30K and **3.9**% R@1 on MSCOCO, illustrating that the additional non-contrastive objective renders extra semantic signals that can be transferred well to fine-grained recognition.

4.2. Representation Learning

We conduct several evaluation protocols to benchmark representation quality under different objectives for both visual (Sec. 4.2.1) and textual (Sec. 4.2.2) modality.

4.2.1 Visual Representation Learning

Linear probing. We evaluate the quality of visual representation via linear probing protocol, where a linear head is fine-tuned on top of the frozen backbones. We follow the same setup as [20] on the same 27 datasets as standard zero-shot classification. Specifically, we use SGD without momentum as the optimizer, no weight decay, and a total epoch of 100 for all evaluation datasets. We use `[CLS]` token for classification. Following [82], we sweep a set of different learning rates by adding multiple classification heads over a shared frozen backbone, each with its own optimizer and scheduler. We report the best result across different heads. As shown in Tab. 5, we observe that nCLIP generally leads to comparable linear probing accuracy compared with CLIP, suggesting that non-contrastive objectives are able to derive semantically meaningful embedding spaces. nCLIP achieves 2.5% higher performance compared to CLIP on ImageNet but relatively lower on average. Beyond that, when combining two objectives together, xCLIP achieves **2.7**% higher performance on ImageNet and **1.5**% higher performance on average across 27 different classification tasks compared to CLIP.

Fine-tuning & semi-supervised learning. We fine-tune the entire network under the full-data regime (100%) and

Model	Flickr30K						MSCOCO					
	I→T			T→I			I→T			T→I		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
CLIP	73.8	93.3	97.1	52.9	79.3	87.7	59.4	85.2	91.9	34.3	62.7	74.0
nCLIP	26.2	53.3	67.9	20.1	48.5	61.1	21.4	47.9	61.6	13.4	37.1	50.3
xCLIP	**77.5**	**95.6**	**97.7**	**57.3**	**82.9**	**89.0**	**63.3**	**87.5**	**94.1**	**38.4**	**66.0**	**76.7**

Table 4. **Zero-shot image-to-text retrieval.** We report R@1, R@5, and R@10 in both image-to-text (I→T) and text-to-image (T→I) settings with ViT-B/16 pre-trained on IT35M. xCLIP consistently improves CLIP in retrieval.

Model	Food101	CIFAR10	CIFAR100	Birdsnap	SUN397	Cars	Aircraft	VOC2007	DTD	Pets	Caltech101	Flowers	MNIST	FER2013	STL10	EuroSAT	RESISC45	GTSRB	KITTI	Country211	PCAM	UCF101	Kinetics700	CLEVR	HatefulMemes	SST	**ImageNet**	**Average**
CLIP	83.3	92.0	75.6	54.4	78.8	60.5	33.8	91.1	72.4	79.0	90.0	96.1	96.2	56.5	97.1	94.8	90.3	77.3	**69.7**	23.2	83.4	78.9	53.9	**50.3**	55.9	**55.8**	71.4	72.7
nCLIP	83.8	93.2	76.3	**66.8**	79.3	49.3	28.0	90.1	70.6	70.2	90.5	95.0	95.4	54.7	97.4	91.9	88.2	75.7	64.0	22.7	**84.2**	79.9	51.0	44.6	55.6	55.2	73.9	71.4
xCLIP	**85.3**	**93.4**	**77.8**	58.4	**80.7**	**62.3**	**36.7**	**92.3**	**74.0**	**81.5**	**91.6**	**97.0**	**96.8**	**58.5**	**98.1**	**95.3**	**91.2**	**80.4**	69.3	**26.3**	83.5	**81.1**	**56.0**	49.8	**58.5**	54.8	**74.1**	**74.2**

Table 5. **Linear probing.** We report on a variety of classification benchmarks with ViT-B/16 pre-trained on IT35M. CLIP achieves consistent performance gains compared to CLIP in a wide range of classification datasets.

Model	sm. sup.		ft.
	1%	10%	100%
CLIP	57.8	72.4	82.4
nCLIP	55.0	72.2	82.4
xCLIP	**59.5**	**73.4**	**82.7**

Table 6. **Fine-tuning (ft.) and semi-supervised learning (sm. sup.).** Percentage shows sampling ratio for end-to-end fine-tuning. We report top-1 accuracy on the ImageNet-1K validation set. nCLIP performs comparatively with CLIP. xCLIP induces consistent gains in all sampling ratios, especially for semi-supervised learning.

Model	STS12	STS13	STS14	STS15	STS16	STS-B	SICK-R	**Average**
CLIP	**52.3**	54.0	53.9	69.9	58.6	**67.0**	69.3	60.7
nCLIP	51.3	**62.7**	**56.9**	**71.9**	**65.8**	66.9	64.1	**62.8**
xCLIP	48.7	55.6	55.6	69.7	59.5	66.3	**70.2**	60.8

Table 7. **Sentence embedding performance on semantic textual similarity.** We report on a variety of NLP understanding benchmarks. xCLIP demonstrates the best capability for various STS tasks. xCLIP has comparable performance with CLIP.

the semi-supervised learning protocol (1% and 10%) on ImageNet-1K [24]. Results are shown in Tab. 6. For CLIP and xCLIP, we find that fine-tuning with the projection head trained under the contrastive objective yields consistent performance gain, especially under a low-data regime. Specifically, we take the text features of the label names as the classifier's initialized parameters. We opt for this setup by default. We highlight that while nCLIP performs worse than CLIP in terms of zero-shot transfer, the non-contrastive objective also serves as a strong baseline to learn representation considering the close gap between CLIP and nCLIP in terms of fine-tuning accuracy. xCLIP achieves an **1.7**%, **1.0**%, and **0.3**% performance gain compared with CLIP under three data ratios of 1%, 10%, and 100%, respectively.

Mask probing. To evaluate how well the visual model is capable of deriving explicit scene layout and object boundaries, we conduct mask probing analysis following [15]. For each attention head from the last layer, we extract the attention map with `[CLS]` token as the query. We then compute the Jaccard similarity $\mathcal{J}$ of each head's attention mask to the ground truth and retain the attention mask with the highest similarity. We conduct experiments on Pascal VOC 2012 [28] dataset. With IT35M, nCLIP demonstrates better quality in the generated mask with 43.7 points of $\mathcal{J}$, while CLIP reaches 41.2 and xCLIP reaches 41.9 points of $\mathcal{J}$. Hence, nCLIP learns better representation for object boundaries compared to CLIP, while xCLIP strikes a balance between nCLIP and CLIP. Detailed results are delayed to Appendix D.1. As a related evaluation, we also study unsupervised segmentation with GroupViT [77] in Appendix D.2.

4.2.2 Textual Representation Learning

We follow the setup as [29] and evaluate on 7 STS tasks: STS 2012–2016 [1–5], STS Benchmark [16], and SICK-Relatedness [34]. We directly take `[EOS]` token without any projection as the input for evaluation, which consistently yields better performance compared to the projected feature for all models. We use cosine distance as the similarity metric. The main goal of sentence embeddings is to cluster semantically similar sentences, and hence, we take STS as one yardstick to benchmark textual representation

Data	Size	Ep.	ZS / LN Accuracy		
			CLIP	nCLIP	xCLIP
①	12M	25	36.8 / 68.5	37.5 / 71.0	**42.4** / **72.2**
④	14M	32	26.5 / 75.1	**31.4** / 77.0	27.1 / **77.2**
①②	22M	32	37.9 / 68.2	34.2 / **72.3**	**43.0** / 71.9
①②③	35M	32	45.7 / 71.4	37.0 / 73.9	**48.8** / **74.1**

Table 8. **Pre-training data scaling & domain.** Ep. denotes training epochs. Data abbreviations are as follows. ①: CC12M. ②: COCO, VG, SBU, and CC3M. ③: YFCC14M. ④: IN21K. Note ①②③ = IT35M.

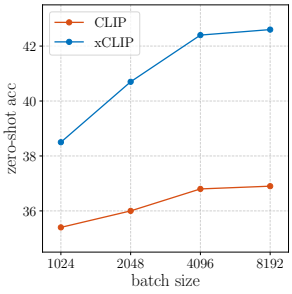

Figure 2. **Batch size scaling.** xCLIP performs better than CLIP with a small batch size (*i.e.*, 1024).

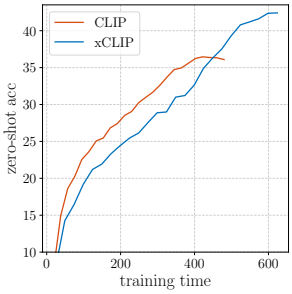

Figure 3. **Training time.** xCLIP adds ∼1.3x training time cost and performs better with equal seen images.

learning. Results are shown in Tab. 7. nCLIP shows better textual semantics compared to CLIP, with **2.1**% higher in terms of the average score with IT35M. The conclusion can also be visually extrapolated from the t-SNE visualization of the word embedding, as detailed in Appendix G. We note that xCLIP performs comparatively with CLIP given that STS is evaluated on `[EOS]` without projection, thus the backbone representation can be similar if dominant by the contrastive objective.

4.3. Properties

We analyze several crucial properties to demonstrate the generalizability and effectiveness of our method in practical usage. Experiments are conducted with ViT-B/16 on CC12M for 25 epochs by default.

Pre-training data scaling. We study the model's performance with data scaling in Tab. 8. xCLIP exhibits desirable scaling behaviors similar to CLIP. The performance improves as data increases. However, we do not observe such a tendency in nCLIP in terms of zero-shot accuracy. For example, nCLIP achieves 37.5% with ① but only 34.2% with ①②, the trend of which defies those of CLIP's (36.8% *vs.* 37.9%) and xCLIP's (42.4% *vs.* 43.0%). On an important note, both nCLIP and xCLIP showcase desirable scaling behaviors under the linear probing protocol. We hypothesize that the performance of zero-shot classification depends on pre-training data quality with the non-contrastive objective, while strong representation can always be ensured before the projection.

Pre-training with tagging data. We consider using *tagging data*, *e.g.*, ImageNet-21K [24], as pre-training data to validate method's generalizability. To do that, we take label names as annotation texts by concatenating them with a sampled prompt, similar to [57, 80]. As shown in Tab. 8, nCLIP shows superiority to CLIP when pre-trained with IN21K by performance gain of **4.9**% and **1.9**% in terms of zero-shot and linear probing accuracy, respectively. Compared to nCLIP, we instead notice a performance drop of 4.3% for zero-shot accuracy when two objectives multi-tasked, indicating that the contrastive objective stymies the quality of projected probability under tagging data. Note, that an **77.2**% linear probing accuracy achieved by xCLIP, 2.1% higher than CLIP, is decently strong compared to 79.0% achieved by the state of the art in SSL, iBOT [82], pre-trained using the same data with 80 epochs.

Batch size scaling. We investigate the scaling behavior of CLIP and xCLIP with different pre-training batch sizes. As shown in Fig. 2, xCLIP performs stronger with smaller batch sizes than CLIP pre-trained even with larger batch sizes. For example, xCLIP pre-trained with a batch size of 1024 achieves a zero-shot accuracy of 38.5% (*vs.* CLIP's 36.8% with a batch size of 4096). However, we still observe a substantial performance drop as batch size decreases, the behavior of which is different from self-supervised models pre-trained with the non-contrastive objective [14, 15] that are insensitive to batch size. We hypothesize that a large batch size is necessary to approximate $\mathbb{E}(p)$ with $\frac{1}{B}\sum_{i=1}^{B} \boldsymbol{p}_i$ in estimating $\mathcal{L}_{\text{HE}}$, especially when two probabilities are derived from drastically different modalities.

Computation efficiency. To demonstrate methods' computation efficiency, we show the FLOPs and GPU memory consumption of xCLIP compared to CLIP. xCLIP enforces bearable additional computation cost on top of CLIP with only 1.4% extra FLOPs and 27% extra GPU memory. As shown in Fig. 3, xCLIP adds ∼1.3x extra time cost but with a 6.0% performance gain in terms of zero-shot accuracy.

4.4. Ablation Study

We show in this section the crucial composing factors of nCLIP and xCLIP. The ablations in the first column (Tabs. 9a to 9c) are conducted with $\mathcal{L}_{\text{nCLIP}}$ only. The ablations in the second column (Tabs. 9d to 9f) are conducted with full loss $\mathcal{L}_{\text{xCLIP}}$. Experiments are conducted with ViT-B/16 on CC12M for 25 epochs. The default settings are highlighted in cyan.

λ_1	λ_2	ZS	LN
0	1	Nan	
0.3	1	26.8	69.5
0.5	1	23.7	67.2
0.2	1.2	35.3	69.0
0.5	1.5	37.5	71.0

(a) **Coefficient of entropy regularizer.** $\lambda_1+1 = \lambda_2$ generally leads to good performance.

dim	ZS	LN	Mem
8192	31.0	70.6	1.07×
16384	35.1	70.9	1.13×
32768	37.5	71.0	1.27×
65536	35.9	71.0	1.52×

(b) **Projection dimension.** Mem (G) is compared to CLIP with a 512-dim embedding layer.

arch	ZS	LN
vanilla	37.5	71.0
w/ bottleneck layer	31.2	69.0
w/ l2-norm layer	35.4	70.7
w/o last BN layer	34.3	69.9

(c) **Projection architecture** of non-contrastive branch.

#SL / #TL	ZS	LN
0 / 1	42.4	72.2
0 / 2	41.3	71.3
1 / 2	42.4	71.9
2 / 3	42.5	72.3

(d) **Shared projection layers.** #SL denotes shared hidden layers. #TL denotes total hidden layers.

$\lambda_{\text{CLIP}} : \lambda_{\text{nCLIP}}$	ZS	LN
1.0	40.2	66.5
0.5	41.7	70.3
0.2	42.4	72.2
0.1	41.8	72.6

(e) **Loss ratio** between $\mathcal{L}_{\text{CLIP}}$ and $\mathcal{L}_{\text{nCLIP}}$.

technique	ZS	LN
vanilla	42.4	72.2
shared space (Eq. (3))	29.4	68.9
debiased sampling	42.2	72.4
warm-up on $\mathcal{L}_{\text{nCLIP}}$	41.7	71.9

(f) **Training objective and technique** for multi-tasking.

Table 9. **Ablation study** with ViT-B/16 on ImageNet-1K validation set. We report zero-shot (ZS) and linear probing (LN) accuracy (%).

Entropy regularizer. We study the effects of coefficient for entropy regularizers λ_1 and λ_2 in Tab. 9a. The pre-training is unstable without regularizers or with only maximization of the entropy of the mean ($\mathcal{L}_{\text{HE}}$). Specifically, the model will collapse to constant uniform distribution and outputs the same uniform probability distribution despite different inputs. The additional minimization of the mean of the entropy ($\mathcal{L}_{\text{EH}}$) stabilizes training, but incurs dimensional collapse, eroding the performances (26.8% and 23.7% *vs.* 37.5% in terms of zero-shot accuracy). Simultaneously adjusting λ_2 with constraints $\lambda_1 + 1 = \lambda_2$ and $\lambda_1 = 0.5$ mitigates the problem and yields the optimal performance, with a 37.5% zero-shot and 71.0% linear-probing accuracy. Details are shown in Appendix B.1.

Projection head. We study the design of two projection heads for $\mathcal{L}_{\text{CLIP}}$ and $\mathcal{L}_{\text{nCLIP}}$. As shown in Tab. 9b, large projection dimension matters for good performance. We opt for a projection dimension of 32768 balancing the performance and speed. As shown in Tab. 9c, the last BN layer is crucial and leads to more stable optimization in our experiments. We draw on the idea of prototypes [9, 15] by introducing bottleneck and l2-norm layers, while they do not yield a performance gain. We further study whether two projections can share intermediate computation in Tab. 9d. While sharing middle hidden layers with slightly deeper projection heads performs better, we opt for 2-layer MLP without shared layers for its simplicity.

Optimization. We study the effect of hyper-parameters in optimization. In Tab. 9e, we study the optimal loss ratio between $\mathcal{L}_{\text{CLIP}}$ and $\mathcal{L}_{\text{nCLIP}}$. We further study whether it's the best route for two objectives to be optimized in separate latent spaces via multi-tasking. Specifically, we consider a bespoke objective Eq. (3) taking account of probability estimation and negative samples simultaneously in a shared latent space. Detailed formulation are shown in Appendix B.2. Optimizing in one latent space yields sub-optimal results with a nearly 13.0% drop in zero-shot accuracy and 3.3% in linear probing accuracy, suggesting two objectives are intrinsically contradictory. We consider debiased sampling, where each batch is sampled from a single data source, since their semantics may be closer and thus more suitable for optimization. This yields similar results. We also add $\mathcal{L}_{\text{nCLIP}}$ after warm-up epochs with the linear scheduler to its base scale, which erodes the performance.

5. Conclusion

Aligning images and texts is of overriding significance for vision-language understanding. To conquer the systematic insufficiency of the contrastive objective for acquiring semantics and tackling loose correlations between noisy image-text pairs, we explore the non-contrastive objective for language-image pre-training and unravel its properties. Empirical evidence reveals that the non-contrastive objective induces models to perform favorably in representation learning yet poorly in zero-shot transfer. Observing the distinct mechanisms of the two objectives, we further seek synergy between the two, and introduce a simple multi-tasking framework, xCLIP, that enjoys the best of both worlds: nCLIP aids CLIP mining semantics while CLIP inherits intrinsic strengths for zero-shot recognition. The consistent performance gain of xCLIP over CLIP on a wide variety of downstream tasks consolidates our findings. As potential future work, we may continue to scale up the data size (*e.g.*, LAION400M [64]) as well as the model size (*e.g.*, ViT-L [26]) to verify whether the scaling law applies and the performance improvement endures.

References

[1] Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Inigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea, et al. Semeval-2015 task 2: Semantic textual similarity, english, spanish and pilot on interpretability. In *SemEval*, 2015. 6

[2] Eneko Agirre, Carmen Banea, Claire Cardie, Daniel M Cer, Mona T Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Rada Mihalcea, German Rigau, and Janyce Wiebe. Semeval-2014 task 10: Multilingual semantic textual similarity. In *SemEval*, 2014. 6

[3] Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez Agirre, Rada Mihalcea, German Rigau Claramunt, and Janyce Wiebe. Semeval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *SemEval*, 2016. 6

[4] Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. Semeval-2012 task 6: A pilot on semantic textual similarity. In *SemEval*, 2012. 6

[5] Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. * sem 2013 shared task: Semantic textual similarity. In *SemEval*, 2013. 6

[6] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*, 2022. 2

[7] Elad Amrani and Alex Bronstein. Self-supervised classification network. In *ECCV*, 2022. 3

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

[9] Mahmoud Assran, Mathilde Caron, Ishan Misra, Piotr Bojanowski, Armand Joulin, Nicolas Ballas, and Michael Rabbat. Semi-supervised learning of visual features by non-parametrically predicting view assignments with support samples. In *ICCV*, 2021. 3, 8

[10] 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 *NeurIPS*, 2019. 4

[11] Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. In *ICLR*, 2022. 3

[12] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *NeurIPS*, 2020. 1

[13] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *ECCV*, 2018. 2, 3

[14] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *NeurIPS*, 2020. 2, 3, 7

[15] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021. 2, 3, 6, 7, 8

[16] Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation. In *SemEval*, 2017. 6

[17] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *CVPR*, 2021. 4

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

[19] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *CVPR*, 2021. 3

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

[21] Yanbei Chen, Yongqin Xian, Sophia Koepke, Ying Shan, and Zeynep Akata. Distilling audio-visual knowledge by compositional contrastive learning. In *CVPR*, 2021. 2

[22] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *ECCV*, 2020. 1, 2

[23] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yan-Tao Zheng. Nus-wide: A real-world web image database from national university of singapore. In *CIVR*, 2009. 5

[24] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, 2009. 2, 4, 6, 7

[25] Karan Desai and Justin Johnson. Virtex: Learning visual representations from textual annotations. In *CVPR*, 2021. 1

[26] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 4, 8

[27] Aleksandr Ermolov, Aliaksandr Siarohin, Enver Sangineto, and Nicu Sebe. Whitening for self-supervised representation learning. In *ICML*, 2021. 3

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

[29] Tianyu Gao, Xingcheng Yao, and Danqi Chen. SimCSE: Simple contrastive learning of sentence embeddings. In *EMNLP*, 2021. 2, 6

[30] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *ICLR*, 2019. 4

[31] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. In *NeurIPS*, 2020. 3

[32] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. In *CVPR*, 2022. 1

[33] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020. 2

[34] Matthew Henderson, Rami Al-Rfou, Brian Strope, Yun-Hsuan Sung, László Lukács, Ruiqi Guo, Sanjiv Kumar, Balint Miklos, and Ray Kurzweil. Efficient natural language response suggestion for smart reply. *arXiv preprint arXiv:1705.00652*, 2017. 6

[35] 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 *ICCV*, 2021. 4

[36] Dan Hendrycks and Kevin Gimpel. Bridging nonlinearities and stochastic regularizers with gaussian error linear units. In *ICLR*, 2017. 4

[37] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *CVPR*, 2021. 4

[38] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. In *ICLR*, 2019. 2

[39] Dat Huynh and Ehsan Elhamifar. A shared multi-attention framework for multi-label zero-shot learning. In *CVPR*, 2020. 5

[40] D. Huynh and E. Elhamifar. A shared multi-attention framework for multi-label zero-shot learning. In *CVPR*, 2020. 5

[41] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015. 4

[42] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, 2021. 2

[43] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *ICML*, 2021. 2

[44] Thomas Kipf, Elise van der Pol, and Max Welling. Contrastive learning of structured world models. *arXiv preprint arXiv:1911.12247*, 2019. 2

[45] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 2017. 4

[46] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4. *IJCV*, 2020. 5

[47] Janghyeon Lee, Jongsuk Kim, Hyounguk Shon, Bumsoo Kim, Seung Hwan Kim, Honglak Lee, and Junmo Kim. Uniclip: Unified framework for contrastive language-image pre-training. In *NeurIPS*, 2022. 2

[48] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In *NeurIPS*, 2021. 2

[49] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. In *ACL*, 2020. 2

[50] Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. In *ICLR*, 2022. 1, 2

[51] 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 *ECCV*, 2014. 4, 5

[52] Xingbin Liu, Jinghao Zhou, Tao Kong, Xianming Lin, and Rongrong Ji. Exploring target representations for masked autoencoders. *arXiv preprint arXiv:2209.03917*, 2022. 1

[53] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 4

[54] Pranay Manocha, Zeyu Jin, Richard Zhang, and Adam Finkelstein. Cdpam: Contrastive learning for perceptual audio similarity. In *ICASSP*, 2021. 2

[55] Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. Slip: Self-supervision meets language-image pre-training. In *ECCV*, 2022. 2

[56] Vicente Ordonez, Girish Kulkarni, and Tamara Berg. Im2text: Describing images using 1 million captioned photographs. In *NeuIPS*, 2011. 4

[57] 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. 1, 2, 4, 7

[58] 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 *ICCV*, 2015. 2, 5

[59] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 1, 2, 3, 4, 5

[60] 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. 1

[61] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *ICML*, 2021. 1

[62] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *ICML*, 2019. 4

[63] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*, 2015. 2

[64] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. In *NeurIPS Workshops*, 2021. 8

[65] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, 2018. 1, 4

[66] Aravind Srinivas, Michael Laskin, and Pieter Abbeel. Curl: Contrastive unsupervised representations for reinforcement learning. In *ICML*, 2020. 2

[67] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. In *ICLR*, 2020. 2

[68] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *EMNLP*, 2019. 2

[69] 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. In *CACM*, 2016. 1

[70] Yael Vinker, Ehsan Pajouheshgar, Jessica Y Bo, Roman Christian Bachmann, Amit Haim Bermano, Daniel Cohen-Or, Amir Zamir, and Ariel Shamir. Clipasso: Semantically-aware object sketching. In *SIGGRAPH*, 2022. 1

[71] Feng Wang, Tao Kong, Rufeng Zhang, Huaping Liu, and Hang Li. Self-supervised learning by estimating twin class distributions. *arXiv preprint arXiv:2110.07402*, 2021. 3

[72] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. In *NeurIPS*, 2019. 4

[73] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. In *ICLR*, 2022. 2

[74] Bichen Wu, Ruizhe Cheng, Peizhao Zhang, Tianren Gao, Joseph E. Gonzalez, and Peter Vajda. Data efficient language-supervised zero-shot recognition with optimal transport distillation. In *ICLR*, 2022. 1, 2

[75] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *CVPR*, 2018. 2

[76] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *CVPR*, 2020. 4

[77] Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. Groupvit: Semantic segmentation emerges from text supervision. In *CVPR*, 2022. 6

[78] Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Bin Xiao, Ce Liu, Lu Yuan, and Jianfeng Gao. Unified contrastive learning in image-text-label space. In *CVPR*, 2022. 1

[79] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. Filip: Fine-grained interactive language-image pre-training. In *ICLR*, 2022. 2

[80] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022. 1, 2, 4, 7

[81] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *ICML*, 2021. 3

[82] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. In *ICLR*, 2022. 5, 7