iCLIP: Bridging Image Classification and Contrastive Language-Image Pre-training for Visual Recognition

Yixuan Wei[1,2] , Yue Cao[2*], Zheng Zhang[2], Houwen Peng[2], Zhuliang Yao[1,2], Zhenda Xie[1,2]
Han Hu[2], Baining Guo[2]
[1]Tsinghua University [2]Microsoft Research Asia

Abstract

This paper presents a method that effectively combines two prevalent visual recognition methods, i.e., image classification and contrastive language-image pre-training, dubbed iCLIP. Instead of naïve multi-task learning that use two separate heads for each task, we fuse the two tasks in a deep fashion that adapts the image classification to share the same formula and the same model weights with the language-image pre-training. To further bridge these two tasks, we propose to enhance the category names in image classification tasks using external knowledge, such as their descriptions in dictionaries. Extensive experiments show that the proposed method combines the advantages of two tasks well: the strong discrimination ability in image classification tasks due to the clean category labels, and the good zero-shot ability in CLIP tasks ascribed to the richer semantics in the text descriptions. In particular, it reaches 82.9% top-1 accuracy on IN-1K, and meanwhile surpasses CLIP by 1.8%, with similar model size, on zero-shot recognition of Kornblith 12-dataset benchmark. The code and models are publicly available at https://github.com/weiyx16/iCLIP.

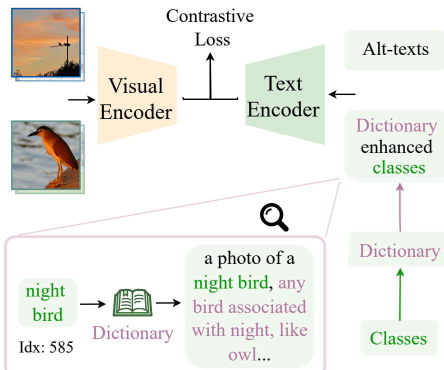

Figure 1. An illustration of the proposed iCLIP framework. The iCLIP framework can take two types of annotations for training: classes and alt-texts. It converts the conventional image classification formula to share the same text encoder and the same cosine classifier as that used in the contrastive language-image pre-training (CLIP). It also uses a dictionary-enhanced approach to enrich the original class names in the image classification problem with external information involved in dictionaries. The deep fusion and knowledge-enriched classes both greatly improve the performance compared to naïve multi-task learning or performing one of the two tasks alone.

1. Introduction

Image classification is a classic visual problem whose goal is to classify images into a fixed set of pre-defined categories. For example, the widely used ImageNet dataset [8] carefully annotated 14 million images and categorize them into 21,841 categories chosen from the WordNet [36]. For image classification, each category provides a clear taxonomy that groups images of the same category together and separates images from different categories, and thus endows the learnt representation with strong discriminant ability. However, this classification ability is limited to a fixed set of categories [8, 29, 51].

Recently, the method that learns to contrast image-text pairs, known as contrastive language-image pre-training (abbr. CLIP), has well made up such shortage of the conventional image classification methods to achieve strong zero-shot recognition ability [24, 44]. These methods employ a contrastive learning framework, where images and their corresponding alt-texts are treated as positive pairs, while images with all other alt-texts are treated as negative pairs. Thanks to the rich semantics involved in the alt-texts, the images can be weakly connected to almost arbitrary categories that already appear in the alt-texts, resulting in its zero-shot ability. A drawback is that the image-text pairs are usually crawled from the internet without human labeling, leading to their noisy and ambiguous nature. Thus the learnt representations are often not conceptual compact, and may lack certain discriminative ability.

*Corresponding Author. The work is done when Yixuan Wei, Zhuliang Yao, and Zhenda Xie are interns at Microsoft Research Asia.

This paper explores how to effectively combine these two powerful visual recognition and representation learning methods, to take advantages of both methods and data sources while relieving their shortages. We first try a naïve multi-task learning framework that applies the original head networks of the two tasks on top of a shared visual encoder, and jointly learn the network with separate losses of the two tasks. This naïve multi-task learning approach has been able to benefit each individual tasks, but the effect is marginal. We thus seek to fuse the two tasks more deeply, so that the advantages of the two tasks can be more effectively joined for better visual recognition, as well as for better transferable representations.

To this end, our first technique is to deeply unify the formulations of image classification and CLIP learning. By examining their formulations, we found there are two main differences: 1) Different classification losses. Image classification tasks typically use a linear classification loss which has better fitting ability due to the non-normalized nature, while the CLIP-based methods adopt a cosine classifier which has better transferability for new domains and categories [2, 6, 9, 18, 38, 57]. 2) Different parameterization methods for classifier weights. Image classification tasks usually directly optimize the parametric classification weights without a need to process text semantics in class names. The CLIP method can be regarded as generating classifier weights through a text encoder and learns the text encoder instead. The text-encoder-based classifier allows sharing between alt-texts as well as modeling their relationships, which enables the ability to tackle any classes.

Although the linear classifier and direct classifier weight parameterization have been common practice in image classification for many years, it is interesting to find that changing the old formulation as that in the CLIP approach has almost no performance degradation for pure image classification problems. This indicates that we can directly adapt the image classification formulation to the cosine classifier and the text encoder parameterization used by CLIP, with almost no loss. This also allows us to further share the text encoder for both class names and alt-texts. Our experiments show that this deep fusion approach performs much better than the naïve multi-task method for both in-domain/zero-shot classification and multi-modal retrieval tasks learning (see 3).

Another gap between the image classification and CLIP lies in the different text richness. Class names are usually in short, i.e., one or a few words, and sometimes are even ambiguous and polysemous in referring to specific semantics, for example, "*night bird*" can represents either "*owl*" or "*nightingale*". On the contrary, alt-texts in CLIP are usually full sentences containing rich information. To further bridge the gap between the image classification and CLIP, we propose a second technique that leverages the *knowledge base* to enhance the original class names, such as the explanations in dictionaries. In our implementation, knowledge is simply encoded as a prefix/suffix prompt, as illustrated in Fig 1. Although simple, dictionary enhanced method shows to maintain the accuracy for pure image classification problem (see Table 1), while greatly improve the zero-shot and multi-modal retrieval performance as shown in Table 2 and 3. Note the process is just like human beings who learn new words or concepts through both real examples and explanations in dictionaries.

By these techniques, we present a framework that deeply fuses the two important tasks of image classification and contrastive language-image pre-training, dubbed iCLIP. Extensive experiments using different combinations of image classification and image-text pair datasets show that the iCLIP method can take advantages of both the discriminative power of image classification tasks and the zero-shot ability in CLIP-like tasks, and perform significantly better than conducting each task alone or the naïve multi-task learning in both the in-domain/zero-shot classification and multi-modal retrieval problems. The iCLIP method also shows that learning a stronger transferable representation than using each of the two tasks alone, verified on a variety of downstream tasks, including ADE20K semantic segmentation [68], LVIS long-tail detection [17], and video action recognition [26], as well as different evaluation settings of few-shot and fine-tuning. Our contributions are summarized as follows:

- We combined two important vision tasks of image classification and contrastive language-image pre-training into a single framework.
- We found that the original image classification formulation can be adapted to CLIP approach with almost no performance degradation. With this finding, we present a deep fusion approach in which the two tasks share the same text encoder and the same classifier type, whose effectiveness is extensively verified on benchmarks.
- We proposed a simple yet effective method to introduce knowledge bases into image classification, addressing the ambiguous and polysemous issue of the originally short image names as well as further bridges the gap between classes and alt-texts. It also provides the first showcase of applying knowledge bases into computer vision problems.

2. Related Work

Supervised visual classification. Classification is almost ubiquitous for visual understanding tasks of various recognition granularity, e.g., image-level classification [12, 20, 28, 33, 49, 52, 58], object-level classification in object detection [3, 15, 19, 45], pixel-level classification in semantic/instance segmentation [5, 35, 63], and video-level action classification [4, 13, 34, 43, 54]. In these tasks, the data is manually annotated to a fixed set of classes, e.g., the 1,000-class ImageNet-1K dataset [8], the 80-class COCO detection dataset [31], the 150-class ADE20K segmentation dataset [68], etc. Among these classification tasks, the image-level classification is particularly important, which has greatly advances the success of deep learning in computer vision, thanks to its high quality and transferable discriminative representations.

The supervised visual classification is generally performed as a K-way classification problem without considering the text semantics of the class names. The most common classifier is the linear classifier, where the classifier vector of each category is parameterized as model weights and is directly learnt through optimization [28].

Contrastive language-image pre-training. Pioneered by CLIP [44] and Align [24], the contrastive language-image pre-training is now attracting more and more attention due to its strong zero-shot transfer capacity. These methods learn a network to pair an image and its associated alt-text, in which the image-text pairs are crawled from the Internet. With web-scale alt-text, it is possible to cover almost all classes, and these methods do show to perform very well for zero-shot recognition. In their frameworks, the images and texts are embedded using two separate encoders, and the output representations of the images and alt-texts are contrasted according to the positive and negative pairs.

While prior to CLIP and Align, there have been a few early works leveraging alt-text or text encoders for image recognition [10,14,16,25,41,46,67]. More follow-up works appeared after CLIP and Align, including Filip [62], DeClip [30], BASIC [42], LiT [66], LiMoE [39], TCL [60], and so on. A drawback of these method is that the image-text pairs are usually noisy without human labeling, leading to the learned representations are not conceptual compact, lacking strong discrimination ability.

Introducing knowledge into AI systems. Our approach is also related to the expert systems in 1980s which heavily rely on a knowledge base for reasoning [23]. Recently, in natural language process, there also emerges boosting large-scale pretrained models by making use of encyclopedic [1, 55] and commonsense knowledge [50]. However, in computer vision, the knowledge bases is not well explored. We hope our findings can encourage more attention to incorporate human knowledge into current vision systems.

Combination of representation learning. Regarding individual strengths of different representation learning approaches, there have been several works trying to combine different representation learning approaches so as to take advantages of individuals' strength. For example, SLIP [37] combines CLIP learning with a self-supervised contrastive learning approach. CoCa [65] combines the CLIP target with an image caption task, in hope to perform well for both understanding and generation problems. MaskCLIP [11] combines CLIP with masked image modeling based self-supervised learning. In contrast, our work also aims to effectively combine different representation learning approaches so as to take both advantages, specifically, the image classification and CLIP.

Relationship to UniCL [61] Concurrent to our work, there is another work named UniCL [61] which also combines image classification with language-image pre-training. We hope the consistent knowledge will help the community in learning more powerful representations. Also note that there are two main differences comparing our framework to the UniCL framework [61]: 1) We involve all negative classifiers in training the supervised classification, while UniCL only involve negatives in a same batch. To make feasible all negative classifiers, we propose a GPU-distributed implementation that distributes the classifiers evenly into different GPUs. Our implementations show to have better in-domain accuracy compared to UniCL when the category number is as large as tens of thousands (76.3% vs. 70.5% as shown in Tab. 4). 2) We introduce a new dictionary enhanced approach to convert the class names with rich semantical text, which shows to be very beneficial for zero-shot image classification and multi-modal retrieval (see Tab. 2).

3. Method

In this section, we first review existing methods on image classification and contrastive language-image pre-training tasks. Then, we propose a unified framework to bridge the two tasks in a deep fusion fashion. Finally, we introduce dictionary-enhanced category descriptions to further align the two taks on input label space.

3.1. Preliminaries

Image Classification. Given a set of <image, category label> pairs, *i.e.*, $\mathcal{D}^c = \{(I_i, C_i)\}_{i=1}^{|\mathcal{D}^c|}$, image classification task targets to predict the category label of a given image, through a visual encoder f_v, and a parametric category classifier h_c, illustrated in Fig. 2 (b). The parameters of h_c is a matrix $W \in \mathcal{R}^{N\times H}$, where N is the number of categories and H is the dimension of visual embeddings. The visual encoder f_v transforms each raw image I_i to an embedding $v_i = f_v(I_i)$, while the classifier h_c predicts the distribution $P_i \in \mathcal{R}^N$ over all pre-defined categories via an inner product between W and v_i, *i.e.*, $P_i = W \cdot v_i$ (bias term is omitted for simplicity). Finally, a cross entropy is applied on P_i and C_i to calculate training loss, which is formulated as:

$$\mathcal{L} = \frac{-1}{|\mathcal{D}^c|} \sum_{(I_i,C_i)\in\mathcal{D}^c} \log \frac{\exp\left(W_{C_i} \cdot v_i\right)}{\sum_{j=1}^{N} \exp\left(W_j \cdot v_i\right)}, \qquad (1)$$

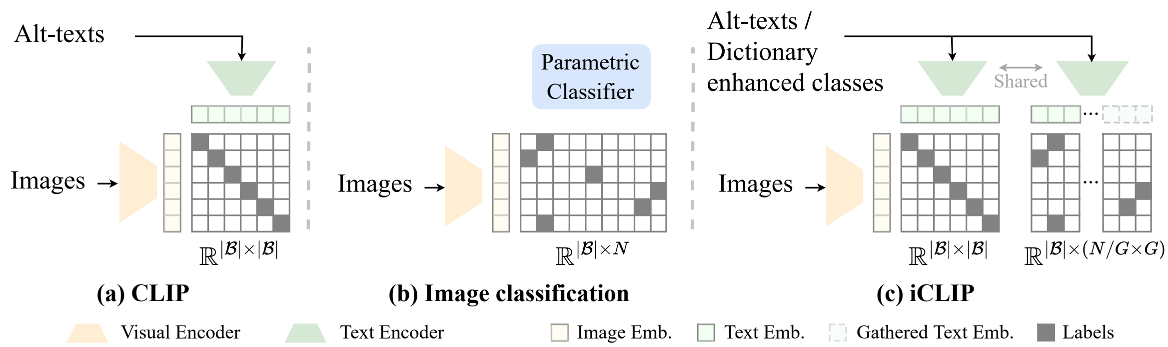

Figure 2. An illustration of iCLIP framework. $\mathcal{B}$ is the batch size, N is the number of categories and G is the number of gpus. iCLIP unifies both contrastive language-image pre-training and classification tasks with shared text and visual encoder, taking alt-texts or dictionary enhanced class names as annotations. To reduce the computation, iCLIP distributes the enhanced class names over all gpus in forward, and gathers the embeddings for similarity calculation.

where W_j is the parametric weight of j-th category.

Contrastive Language-Image Pre-training. Given a set of <image, alt-text> pairs, *i.e.*, $\mathcal{D}^a = \{(I_i, T_i^a)\}_{i=1}^{|\mathcal{D}^a|}$, contrastive language-image pre-training targets to close the distances between paired image and text while enlarging those of unpaired ones, through a visual encoder f_v and a text encoder f_t, shown in Fig. 2 (a). They transform the image I_i and the alt-text T_i^a to feature embeddings v_i and s_i, respectively. A contrastive loss function is applied to shrink the cosine distance of v_i and s_i, which is defined as:

$$\mathcal{L} = \frac{-1}{|\mathcal{D}^a|} \sum_{\substack{(I_i, T_i^a) \\ \in \mathcal{D}^a}} \log \frac{\exp\left(\cos\left(f_t\left(T_i^a\right), v_i\right)/\tau\right)}{\sum\limits_{T_j^a \in \mathcal{T}^a} \exp\left(\cos\left(f_t\left(T_j^a\right), v_i\right)/\tau\right)}, \quad (2)$$

where $\cos(\cdot,\cdot)$ represents the cosine similarity between two embeddings, $\mathcal{T}^a$ is all the alt-texts in a batch including one positive paired alt-text and $|\mathcal{T}^a| - 1$ negative ones, and τ is a temperature hyper-parameter to scale the similarities.

Task differences. Comparing the formations of image classification and language-image pre-training, we can draw three main difference between them. 1) *Training loss functions*. Classification commonly adopts a cross-entropy loss on inner-product similarity, while image-text learning uses InfoNCE loss on cosine similarity. 2) *Classifier types*. Classification adopts a parametric category classifier, while image-text learning uses a text encoder. 3) *Label granularity*. Category names in classification are usually very short, *i.e.*, one or few words, while the captions in image-text pre-training are full sentences containing rich semantics.

3.2. Bridge Image Classification and Contrastive Language-Image Pre-training

To bridge image classification and image-text alignment, we introduce three adaptations to align their training losses, unify the classifier types, and close the label granularity gap. The overall adaption is visualized in Fig. 3.

Classification with Text Encoder. As formulated in Eq. (1), image classification commonly adopts a cross-

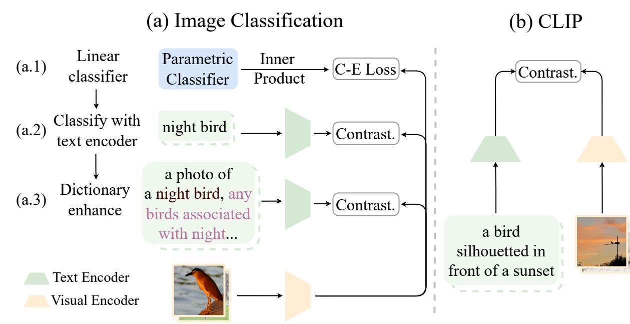

Figure 3. An illustration of our approach to bring image classification (a) to CLIP (b), from the perspective of loss function, classifier types and label granularity. We reformulate the linear classifier (a.1) with a text-encoder-based classifier (a.2), and enhance the class names with a text description from the dictionary (a.3).

entropy loss on top of the inner-product similarity between the visual embedding v_i and the parametric classifier h_c. This formulation is not in line with the InfoNCE loss in Eq. (2), leading to a misalignment between the two paradigms. To address this issue, we adopt a cosine similarity for image classification, instead of the original inner-product similarity in Eq. (1), which formulates a cosine classifier as:

$$\mathcal{L} = \frac{-1}{|\mathcal{D}^c|} \sum_{(I_i, C_i) \in \mathcal{D}^c} \log \frac{\exp\left(\cos\left(W_{C_i}, v_i\right)/\tau\right)}{\sum_{j=1}^{N} \exp\left(\cos\left(W_j, v_i\right)/\tau\right)}. \quad (3)$$

Cosine similarity is a common practice in metric learning [40]. It can smoothly align the supervised image classification with the cross-modal contrastive pre-training in terms of learning objective function, *i.e.*, Eq. (2). Moreover, our experiments demonstrate that this cosine classifier performs on par with the traditional linear classifier (see Tab. 1).

The cosine classifier aligns the training losses of two tasks. However, the annotations, *i.e.*, category labels and captions, are modeled separately by the parametric category classifier h_c and the text encoder f_t. As analyzed in Sec. 4.3, shallowly combining the two tasks with a shared visual encoder f_v and two separate task heads does not fully take advantage of the gold annotations in image classification and rich concepts in textual captions, resulting in a sub-optimal solution with limited transferring capacity.

To tackle this issue, we take label semantics into consideration and propose to utilize the text encoder f_t as a meta classifier for image classification. Formally, we replace the label index C_i with its class name M_i, and generate the classifier weight W on-the-fly through the text encoder f_t, which is shared with image-text pre-training. The new formulation is represented as:

$$\mathcal{L} = \frac{-1}{|\mathcal{D}^c|} \sum_{\substack{(I_i, M_i) \\ \in \mathcal{D}^c}} \log \frac{\exp\left(\cos\left(f_t\left(M_i\right), v_i\right) / \tau\right)}{\sum_{j=1}^{N} \exp\left(\cos\left(f_t\left(M_j\right), v_i\right) / \tau\right)}. \quad (4)$$

In this way, the text encoder f_t is not only used to extract semantics from gold category labels, but also capture textual information from image captions. Both the visual and textual encoders are shared across the two tasks, leading to a deep fusion of the two tasks.

Classification with Dictionary Enhancement. The cosine classifier with text encoder as a meta network has largely unify the two tasks in model training. In this step, we further align them on input label granularity, reducing the disparity between label names (one or few words) and image captions (a complete sentence). Our proposal is to integrate external knowledge into label names. More specifically, for each label names, we introduce detailed descriptions from its corresponding synset in the dictionary WordNet [36] as the external knowledge and create a pseudo sentence as label for each categories. We combine the original class names and their dictionary descriptions to form the enhanced texts as the input to the text encoder. Also, we add a prompt to make the sentence more fluent. The final dictionary-enhanced description for each category is formed as:

$$\mathcal{T}^c = \text{A photo of a } \{\text{NAME}\}_{C_i}, \{\text{DESCRIPTION}\}_{C_i}. \quad (5)$$

Such dictionary-enhanced descriptions have similar label granularity to alt-text, and thus further bring image classification closer to image-text alignment. Moreover, the description introduces more details of each category, being capable of reducing potential misconception. For example, the class "*night bird*" actually includes several kinds of birds, like owl, nightingale, *etc*. Such a category name cannot allow the model to learn precise representations due to the blurry concepts. If we augment the category with more external knowledge, such as "*a photo of a night bird, any bird associated with night: owl, nightingale, nighthawk*", it will help the model learn discriminative representation on distinguishing different concepts (*e.g.*, bird species).

A Unified Framework. The above three steps adapt image classification to image-text alignment from the perspective of training loss, classifier type and annotation granularity, respectively. Towards the final unification, we propose a new framework dubbed iCLIP, as presented in Fig. 2 (c), which bridges Image Classification and Image-Text Alignment with a unified contrastive learning loss formulated as:

$$\mathcal{L} = \frac{-1}{|\mathcal{D}|} \sum_{(I_i, T_i) \in \mathcal{D}} \log \frac{\exp\left(\cos\left(f_t\left(T_i\right), v_i\right) / \tau\right)}{\sum\limits_{T_j \in \mathcal{T}} \exp\left(\cos\left(f_t\left(T_j\right), v_i\right) / \tau\right)}, \quad (6)$$

where $\mathcal{D}$ is a set consisting of the image classification data $\mathcal{D}^c$ and the image-text alignment data $\mathcal{D}^a$, *i.e.*, $\mathcal{D} = \{\mathcal{D}^c, \mathcal{D}^a\}$, while $\mathcal{T}$ indicates a combination of $\mathcal{T}^c$ and $\mathcal{T}^a$, *i.e.*, $\mathcal{T} = \{\mathcal{T}^c, \mathcal{T}^a\}$. Text label T_i is either an image caption T_i^a sampled from $\mathcal{T}^a$ or a dictionary-enhanced description T_i^c sampled from $\mathcal{T}^c$. It is worth noting that, with this unified framework, both the text encoder f_t and the visual encoder f_v are shared across the two tasks, achieving a deep fusion. The proposed unified framework is able to leverage any combination of tag-labeled and caption-labeled image datasets for pre-training. This combination allows the model to learn more discriminative representation, while capturing more visual concepts from the textual description. On the other hand, our iCLIP method is efficient.

Distributed Implementation. In our iCLIP framework, the text embedding of each category is generated by the shared text encoder on-the-fly. This computation is affordable when the number of categories N is not large. However, it will become infeasible if category number scales up to be large, such as 22k categories in ImageNet-22K [8]. To make the iCLIP framework feasible for large-category classification data in practice, we adopt a distribution implementation strategy [6]. Specifically, we distribute all the enhanced class names evenly over G GPUs in forward, and gather the embeddings from each gpu for similarity calculation, reducing the computation cost and saves memory consumption by the text encoder to $1/G$.

Table 1. Ablation on formulation adaptations for image classification task. Models are trained with 100 epochs.

#	Cosine Loss	Text-enc. as Classifier	Enhanced classes	IN-1K
1				80.9
2	✓			81.5
3	✓	✓		81.2
4	✓	✓	✓	81.4

4. Experiment

We verify the effectiveness of the proposed iCLIP framework through the comparisons to single-task baselines and a naïve multi-task learning baseline. The comparisons are conducted in three settings covering different scales of pre-training data. In evaluation, we assess the models on different tasks, including in-domain classification, zero-shot classification, multi-modal retrieval, and downstream tasks.

Table 2. Ablation study conducted on IN-22K [8] and YFCC-14M [53]. Models are pre-trained from scratched with 32 epochs following UniCL [61]. COCO and Flickr stand for MSCOCO [31] and Flickr30K [64]. IR and TR stand for image retrieval and text retrieval.

			Zero-shot classification		Zero-shot retrieval			
#	Training Data	Method	IN-1K	14-dataset avg.	Flickr-IR	Flickr-TR	COCO-IR	COCO-TR
1	YFCC-14M	CLIP [44]	30.1	36.3	21.5	37.9	12.5	21.2
2	YFCC-14M (half) + IN-21K (half)	*iCLIP (w/o Desc.)*	39.4	45.4	27.6	39.1	13.0	20.4
3	YFCC-14M (half) + IN-21K (half)	iCLIP	**45.9**	**49.9**	**31.9**	**49.8**	**15.5**	**27.2**
4	YFCC-14M + IN-21K	*iCLIP (w/o Desc.)*	41.1	49.4	33.4	51.2	16.3	26.5
5	YFCC-14M + IN-21K	iCLIP	**50.9**	**54.4**	**37.1**	**55.7**	**18.5**	**30.7**
6	YFCC-14M + IN-22K	*iCLIP (w/o Desc.)*	76.2	51.6	33.2	48.2	14.4	23.8
7	YFCC-14M + IN-22K	iCLIP	**76.3**	**55.5**	**36.2**	**55.3**	**18.0**	**29.7**

4.1. Experimental Setup

Pre-training data and settings. We consider three different scales of dataset combination for model pre-training.

- *ImageNet-1K [8] and GCC-3M [48].* In this setting, we use ImageNet-1K as the classification data while GCC-3M as the image-text data. We adopt a Swin-T [33] initialized with MoBY [59] as the visual encoder, while for the textual encoder, we use a pretrained RoBERTa-B [32]. We sample half number of images from each dataset in a mini-batch and train the models with a batch size of 128 $\times$ 8 V100 GPUs for 100 epochs. The highest learning rate is 2e-4 with a cosine learning rate schedule and 5 epochs warm-up. Weight decay is set to be 0.01. RandAugment [7] and stochastic depth [21] with a rate of 0.1 are used for visual encoder only.
- *ImageNet-22K [8] and YFCC-14M [53].* We follow UniCL [61] to train all models from scratch with 32 epochs for a fair comparison with it. Swin-T [33] is used as the visual encoder, and a 12-layer transformer with a hidden dimension of 512 same as CLIP [44] is used as the text encoder. A batch size of 512 $\times$ 16 GPUs is adopted. The highest learning rate is selected from 2e-4 and 8e-4. Other regularization is the same as previous, except for a larger weight decay of 0.05. We also conduct experiments using two variants of this setup for a fair and clean comparison with the methods that use one task alone (IC or CLIP): 1) Excluding the 1,000 ImageNet-1K classes in ImageNet-22K dataset (dubbed IN-21K). This setup variant allows us to evaluate the zero-shot accuracy on ImageNet-1K for different methods; 2) Half images of the ImageNet-21K and YFCC-14M are used, such that the dataset size and training iterations are the same as that used in one single task.
- *ImageNet-22K [8] and Laion-400M [47].* For this large-scale pre-training setting, we adopt a Swin-B initialized with MoBY as the visual encoder and a pre-trained RoBERTa-B as the text encoder. We train iCLIP for 100K iters, with a batch size of 192 $\times$ 64 V100 GPUs. In each mini batch, we sample 64 images from IN-22K and 128 images from Laion-400M. The model is trained on classification data for around 30 epochs and on image-text data for around 2 epochs equivalently. The highest learning rate is 1e-3 with a cosine learning rate schedule and a warm-up for 16.7K iters. Weight decay is set to 0.05 and drop depth rate is set to 0.2.

Evaluation datasets and settings. During evaluation, we assess the models considering five different settings.

- *Zero-shot classification*. We evaluate the concept coverage and generalization ability of the models on three datasets: 1) ImageNet-1K variants, including IN-1K [8], and IN-Sketch (IN-S) [56]. Top-1 accuracy is reported; 2) the widely-used Kornblith 12-dataset benchmark [27]; 3) 14 datasets used in UniCL [61]. For 2) and 3), averaged accuracy is reported.
- *Zero-shot multi-modal retrieval*. Flickr30K [64] (1K test set) and MSCOCO [31] (5K test set) are used to evaluate the alignment between image and text modalities. We report the Top-1 recall on both image retrieval (IR) and text retrieval (TR).
- *In-domain classification*. ImageNet-1K data is included in some of our pre-training setups, so we conduct in-domain evaluation on ImageNet-1K in these cases. The Top-1 accuracy is reported.
- *Few-shot classification*. Following CLIP [44], we also evaluate the models on few-shot classification task using Kornblith 12-dataset with a frozen visual encoder. Averaged accuracy is reported.
- *Fine-tuning on downstream tasks*. To validate the generalization ability of iCLIP, the models are fine-tuned and compared on semantic segmentation [68], long-tail detection [17], and video action recognition [26]. We report val mIoU, bbox mAP and Top-1 accuracy, respectively. The detailed settings can be found in the *supplementary material*.

4.2. Experiments on IN-1K [8] and CC3M [48]

Formulation adaptations for image classification. Tab. 1 ablates the effect of adapting the common image classification to that used in iCLIP, including both cosine loss, the

Table 3. Ablation conducted on IN-1K [8] and GCC-3M [48] combined data. For the models only using IN-1K, we train them for 100 epochs. For the models only using GCC-3M, we train them with the same iterations and batch size as the ones used in IN-1K.

		12-dataset	ImageNet-related	
#	Method	avg.	IN-1K	IN-S
1	*Sup-only*	-	**80.9**	29.4
2	*VL-only*	31.4	32.4	18.3
3	*Naïve multi-task*	35.1	80.6	38.3
4	*iCLIP (w/o Desc.)*	37.7	80.5	38.6
5	iCLIP	**39.1**	80.4	**38.7**

Table 4. Comparison with UniCL. Models are pre-trained from scratched with 32 epochs, following UniCL [61].

#	Training Data	Method	IN-1K	14-dataset avg.
1	YFCC + IN-21K (half)	UniCL [61]	36.4	45.5
2	YFCC + IN-21K (half)	iCLIP	**45.9**	**49.9**
3	YFCC + IN-21K	UniCL [61]	40.5	49.1
4	YFCC + IN-21K	iCLIP	**50.9**	**54.4**
5	YFCC + IN-22K	UniCL [61]	70.5	52.4
6	YFCC + IN-22K	iCLIP	**76.3**	**55.5**

text-encoder-based classifier and enhanced class names using ImageNet-1K dataset. It can be seen that the cosine classification loss gets slightly better performance than the linear one, with a +0.6% gain on IN-1K (see #1 v.s. #2). When further adapting the text-encoder-based classifier (#3) and enhancing class names from dictionaries (#4), it has almost no performance degradation (+0.3% and +0.5% on IN-1K compared to the linear classifier), which allows to further sharing the text encoder with CLIP for tasks unification.

Zero-shot and in-domain classification. With previous adaptions on the image classification formulation, we can further share the text encoder between the two tasks. To ablate the effect of sharing the text encoder, we set a *naïve multi-task* baseline, that combines image classification and CLIP in a shallow fusion, *i.e.*, simply averaging the loss Eq. (1) and Eq. (2). Each has its own head network, *i.e.*, the fully-connected layer W for Eq. (1) and the text encoder f_t for Eq. (2). The best performances of the two heads are reported in Tab. 3. With a shared text encoder across the two tasks, our *iCLIP (w/o Desc.)* outperforms the *naïve multi-task* on Kornblith 12-dataset zero-shot classification by +2.6% in average, while they are comparable on ImageNet-related datasets classification (see #3 v.s. #4). Our iCLIP deeply unifies two tasks, thus better gathering the merits of the two learning protocols. When compared with the supervised softmax classifier baseline, *i.e.*, Eq. (1) *Sup-only*, and the contrastive image-text pre-training baseline, *i.e.*, Eq. (2) *VL-only*, our method is sightly worse than *Sup-only* on IN-1K by 0.4%, while achieves superior performance on other evaluation settings, +6.3% better than *VL-only* method on 12-dataset zero-shot testing and +9.2% better than *Sup-only* method on IN-S (see #4 v.s. #1&#2). Moreover, the dictionary enhancement on class names (#5) can further bring an average of +1.4% improvements on Kornblith 12-dataset, revealing the increased discriminative representation for ambiguous concepts.

4.3. Experiments on IN-22K [8] and YFCC14M [53]

Effects of the unified framework. Here, we further ablate the effect of the unified formulation for deep fusion of the two tasks. In #2, #4 and #6 of Tab. 2, we show the results of our unified framework under three different dataset combination setups. Compared with the CLIP baseline (#1), our iCLIP (#2) earns +8.3% gains on IN-1K zero-shot classification and also +9.1% improvements when evaluated on the 14-dataset. In addition, our iCLIP is better than the CLIP baseline on most cross-modal retrieval benchmarks, while only using half of visual-language data in pre-training.

Effects of dictionary enhancement. Furthermore, we dissect the model to study the contributions of dictionary-enhanced category description. From Tab. 2, we can see that enhancing each class names with informative description from the dictionary brings consistent improvements on both zero-shot classification and zero-shot retrieval under three dataset combination setups (see #3, #5 and #7). In particular, when pre-trained with half images of YFCC-14M and IN-21K (#3), the integrated knowledge contributes +6.5% improvements on IN-1K zero-shot classification, which makes our iCLIP reach 45.9%, being +5.4% better than UniCL method [61] with full images of YFCC-14M and IN-21K (see #3 in Tab. 4). More importantly, the enhanced class names is beneficial to cross-modal retrieval. For example, for image-to-text search, the dictionary-enhanced description can bring 10.7% and 6.8% top-1 recall gains on Flickr30K [64] and MSCOCO [31] respectively, as reported in row 3 of Tab. 2.

Comparison with UniCL [61]. Tab. 4 summaries our comparison to UniCL. The same as UniCL, we evaluate our models on IN-1K and 14 datasets. Under three different dataset combination setups, our iCLIP surpasses UniCL by at least +5% on IN-1K image classification, while reaching 55.5% averaged accuracy on 14 datasets (#6), being +3.1% better than UniCL (#5).

4.4. Experiments on IN-22K and Laion-400M [47]

Zero-shot and in-domain classification. Tab. 5 presents a large scale experiment using the publicly accessible large-scale data: Laion-400M [47] and IN-22K [8]. For *Sup-only*, *i.e.* Eq. (1), we use the released version from Swin [33], which is trained on IN-22K for 90 epochs. For *VL-only*, *i.e.* Eq. (2), we pre-train it on Laion-400M with a similar image numbers (#im). Our method is comparable to *Sup only* on IN-1K, while it gets +17.8% and +8.3% better results than the two baselines on IN-S, demonstrating its robustness to natural distribution shifts. Our iCLIP surpasses

Table 5. Ablation study on IN-22K [8] and Laion-400M [47]. We evaluate the models on ImageNet datasets (IN-1K [8] and IN-S [56]) and zero-shot evaluation on the Kornblith 12-dataset benchmark [27]. Few-shot learning on Kornblith 12-dataset and the fine-tuning on three downstream tasks are conducted to evaluate the transfer capability of iCLIP. ‡ denotes for our reproduction using released checkpoints.

	Visual encoder	Pre-train	ImageNet-related		12-dataset avg.		downstream tasks		
Method	Arch.	length (#im).	IN-1K	IN-S	0-shot	4-shot	ADE20K	LVIS	Kinetics400
CLIP [44]	ViT-B/16	$400M\times32$ eps	68.6	46.6‡	68.8	66.4‡	-	-	-
OpenCLIP [22]	ViT-B/16	$400M\times32$ eps	67.1	52.4‡	70.9‡	-	-	-	-
Sup-only	Swin-Base	$14M\times90$ eps	82.6	42.0	-	67.6	52.1	35.9	82.7
VL-only	Swin-Base	$400M\times3$ eps	61.1	51.5	67.2	73.3	52.0	36.6	82.3
iCLIP	Swin-Base	$400M\times2$ eps+$14M\times30$ eps	**82.9**	**59.8**	70.6	**78.1**	**52.6**	**37.9**	**83.1**

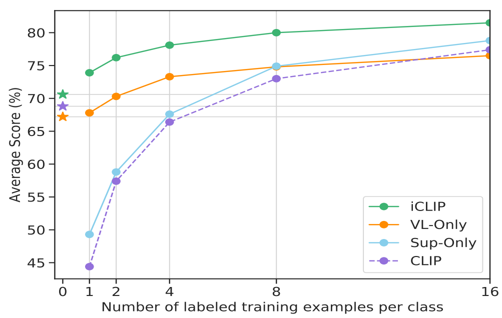

Figure 4. Major comparison with the CLIP-ViT-B/16 of few-shot classification (top-1 accuracy) on the Kornblith 12-dataset. ⋆ denotes the zero-shot performances. Results of CLIP on few-shot classification are reproduced using released model. We run every experiments three times and the averaged results are reported.

OpenCLIP [22], which also uses Laion-400M data for pre-training [47], by more than +15% on IN-1K, mainly due to the pre-training data IN-22K covers the visual concepts in IN-1K. Moreover, when performing zero-shot evaluation on 12 datasets [27], our iCLIP model also achieves non-trivial improvements, *e.g.*, an average of over +3% gains (*VL-only* in Tab. 5). In addition, our iCLIP is comparable to OpenCLIP on 12 datasets in average with fewer training time. More details are elaborated in the *supplementary material*.

Few-shot classification We also conduct experiments in few-shot settings. Following CLIP [44], we freeze the visual encoder and append a linear probe layer for few-shot fine-tuning. We notice that the performance of CLIP [44] in few-shot classification cannot catch up with that of zero-shot classification, unless more than 4 examples per class are given, as presented in Fig. 4 (⋆ *v.s.* -●-). We conjecture the underlying reason is that the number of training samples is too limited to train a randomly initialized classifier. This situation can be alleviated by fine-tuning the pretrained text encoder, instead of the linear probe layer. In this way, text encoder is able to serve as a good initialization for few-shot classification, closing the gap between pretraining and fine-tuning. We evaluate such method on Kornblith 12-dataset benchmark [27] and report the results in Fig. 4.

When only given one example per class, by utilizing text encoder as the classifier, our iCLIP achieve 73.9% on 12-dataset in average, surpassing the original CLIP model by +29.5%. Such one-shot recognition gets +3.3% gains over the zero-shot baseline (⋆ *v.s.* -●-), demonstrating good few-shot transfer ability. When using 16 examples per class, our model still performs superior to CLIP by 4.1%. Compared to supervised-only model and visual-linguistic only model, our unified contrastive learning pretrained model obtains +24.6% and +6.1% better accuracy under one-shot learning setting. Such advantages are kept to 16-shot with +2.7% and +5.0% gains (-●- and -●-).

Fine-tuning on Downstream Tasks We also study the generalization capability of our pre-trained models on downstream tasks, including semantic segmentation, object detection and video recognition. As shown in Tab. 5, compared to *Sup-only*, our iCLIP surpasses it by +0.5%, +2.0%, +0.4% on the three downstream tasks, respectively. We also earn +0.6%, +1.3%, +0.8% gains over *VL-only* baseline. These results reveal that our unified method could learn general visual representations.

5. Conclusion

In this paper, we propose a unified framework dubbed iCLIP to bridge image classification and language-image pre-training. It naturally forces the cross-modal feature learning in a unified space, where the two tasks share the same visual and textual encoders. Extensive experiments demonstrate that iCLIP is effective, and can be generalized to different visual recognition scenarios, including zero-shot, few-shot, and fully-supervised fine-tuning.

Limitations. One limitation of iCLIP is that, despite its competitive performance, the model still relies on human labeled classification data that is not scalable. Besides, our model currently only adopts median-size parameters, which can not fully validate the generation ability to large-scale models. We are interested in exploring this in future work.

References

[1] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. Dbpedia: A nucleus for a web of open data. In *The semantic web*, pages 722–735. Springer, 2007. 3

[2] Yue Cao, Zhenda Xie, Bin Liu, Yutong Lin, Zheng Zhang, and Han Hu. Parametric instance classification for unsupervised visual feature learning. 2020. 2

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020. 3

[4] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 3

[5] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 3

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. pages 1597–1607, 2020. 2, 5

[7] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 702–703, 2020. 6

[8] 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. 1, 3, 5, 6, 7, 8

[9] Jiankang Deng, Jia Guo, Jing Yang, Niannan Xue, Irene Cotsia, and Stefanos P Zafeiriou. ArcFace: Additive angular margin loss for deep face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2021. 2

[10] 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. 3

[11] Xiaoyi Dong, Yinglin Zheng, Jianmin Bao, Ting Zhang, Dongdong Chen, Hao Yang, Ming Zeng, Weiming Zhang, Lu Yuan, Dong Chen, et al. Maskclip: Masked self-distillation advances contrastive language-image pretraining. *arXiv preprint arXiv:2208.12262*, 2022. 3

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 2

[13] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 6202–6211, 2019. 3

[14] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc' Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. 3

[15] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014. 3

[16] Lluis Gomez, Yash Patel, Marçal Rusiñol, Dimosthenis Karatzas, and CV Jawahar. Self-supervised learning of visual features through embedding images into text topic spaces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4230–4239, 2017. 3

[17] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5356–5364, 2019. 2, 6

[18] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. pages 9729–9738, 2020. 2

[19] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 3

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

[21] Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q Weinberger. Deep networks with stochastic depth. In *European conference on computer vision*, pages 646–661. Springer, 2016. 6

[22] 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. 8

[23] P Jackson. Introduction to expert systems. 1 1986. 3

[24] 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. 1, 3

[25] Armand Joulin, Laurens Van Der Maaten, Allan Jabri, and Nicolas Vasilache. Learning visual features from large weakly supervised data. In *European Conference on Computer Vision*, pages 67–84. Springer, 2016. 3

[26] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics hu-

man action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 2, 6

[27] Simon Kornblith, Jonathon Shlens, and Quoc V. Le. Do better imagenet models transfer better? In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2656–2666, 2019. 6, 8

[28] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 2, 3

[29] 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. *International Journal of Computer Vision*, 128(7):1956–1981, 2020. 1

[30] 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 *International Conference on Learning Representations*, 2022. 3

[31] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 3, 6, 7

[32] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 6

[33] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. pages 10012–10022, 2021. 2, 6, 7

[34] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3202–3211, June 2022. 3

[35] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 3

[36] George A. Miller. WordNet: A lexical database for English. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*, 1994. 1, 5

[37] Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. Slip: Self-supervision meets language-image pretraining. In *European Conference on Computer Vision*, pages 529–544. Springer, 2022. 3

[38] Kevin Musgrave, Serge Belongie, and Ser-Nam Lim. A metric learning reality check, 2020. 2

[39] Basil Mustafa, Carlos Riquelme, Joan Puigcerver, Rodolphe Jenatton, and Neil Houlsby. Multimodal contrastive learning with limoe: the language-image mixture of experts. *arXiv preprint arXiv:2206.02770*, 2022. 3

[40] Hieu V Nguyen and Li Bai. Cosine similarity metric learning for face verification. In *Asian conference on computer vision*, pages 709–720. Springer, 2010. 4

[41] Mohammad Norouzi, Tomas Mikolov, Samy Bengio, Yoram Singer, Jonathon Shlens, Andrea Frome, Greg S Corrado, and Jeffrey Dean. Zero-shot learning by convex combination of semantic embeddings. *arXiv preprint arXiv:1312.5650*, 2013. 3

[42] 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. *CoRR*, abs/2111.10050, 2021. 3

[43] Zhaofan Qiu, Ting Yao, and Tao Mei. Learning spatio-temporal representation with pseudo-3d residual networks. In *proceedings of the IEEE International Conference on Computer Vision*, pages 5533–5541, 2017. 3

[44] 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. 1, 3, 6, 8

[45] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015. 3

[46] Mert Bulent Sariyildiz, Julien Perez, and Diane Larlus. Learning visual representations with caption annotations. In *European Conference on Computer Vision (ECCV)*, 2020. 3

[47] 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, 2021. 6, 7, 8

[48] 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 ACL*, 2018. 6, 7

[49] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, May 2015. 2

[50] Robyn Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Thirty-first AAAI conference on artificial intelligence*, 2017. 3

[51] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 1

[52] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015. 2

[53] 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. *Commun. ACM*, 59(2):64–73, jan 2016. 6, 7

[54] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015. 3

[55] Denny Vrandečić. Wikidata: A new platform for collaborative data collection. In *Proceedings of the 21st international conference on world wide web*, pages 1063–1064, 2012. 3

[56] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems*, pages 10506–10518, 2019. 6, 8

[57] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition, 2018. 2

[58] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1492–1500, 2017. 2

[59] Zhenda Xie, Yutong Lin, Zhuliang Yao, Zheng Zhang, Qi Dai, Yue Cao, and Han Hu. Self-supervised learning with swin transformers. *arXiv preprint arXiv:2105.04553*, 2021. 6

[60] Jinyu Yang, Jiali Duan, Son Tran, Yi Xu, Sampath Chanda, Liqun Chen, Belinda Zeng, Trishul Chilimbi, and Junzhou Huang. Vision-language pre-training with triple contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15671–15680, 2022. 3

[61] Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Bin Xiao, Ce Liu, Lu Yuan, and Jianfeng Gao. Unified contrastive learning in image-text-label space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19163–19173, June 2022. 3, 6, 7

[62] 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 *International Conference on Learning Representations*, 2022. 3

[63] Minghao Yin, Zhuliang Yao, Yue Cao, Xiu Li, Zheng Zhang, Stephen Lin, and Han Hu. Disentangled non-local neural networks. In *Proceedings of the European conference on computer vision (ECCV)*, 2020. 3

[64] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. 6, 7

[65] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *Transactions on Machine Learning Research*, 2022. 3

[66] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18123–18133, 2022. 3

[67] 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. 3

[68] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal on Computer Vision*, 2018. 2, 3, 6