FuseTeacher: Modality-fused Encoders are Strong Vision Supervisors

Chen-Wei Xie[1], Siyang Sun[1], Liming Zhao[1],
Pandeng Li[2], Shuailei Ma[3], and Yun Zheng[1]

[1] Alibaba Group
{eniac.xcw,siyang.ssy,lingchen.zlm}@alibaba-inc.com
[2] University of Science and Technology of China
[3] Northeastern University

Abstract. Learning visual representation with image-text datasets attracts a lot of attention in recent years. Existing approaches primarily rely on cross-modality supervision, and incorporate intra-modality supervision if necessary. They overlook the potential benefits of modality-fused supervision. Since modality-fused representation augments the image representation with textual information, we conjecture it is more discriminative and potential to be a strong teacher for visual representation learning. In this paper, we validate this hypothesis by experiments and propose a novel method FuseTeacher that learns visual representation by modality-fused supervision. Specifically, we introduce a fusion encoder that encodes image and text into a fusion representation. This representation can be utilized to supervise the visual representation learning in two distillation ways: (i) Classification Distillation: we cluster image-text pairs into K clusters using the fusion representation and assign each pair a soft cluster assignment, which is served as a pseudo classification label for supervising the image encoder. (ii) Retrieval Distillation: we calculate the similarities between the fusion representation and all text representations in the same batch. By using the similarity distribution as pseudo retrieval similarity between the corresponding image and all texts, we can prevent one-to-one contrastive learning from separating relevant but unpaired pairs. The FuseTeacher is compatible with existing language supervised visual representation learning methods. Experimental results demonstrate that it is able to bring significant improvements and achieves state-of-the-art methods on various datasets. Code, datasets and pre-trained models are available at `https://github.com/Eniac-Xie/FuseTeacher`.

1 Introduction

Visual representation learning with language supervision [30,32,54] has recently achieved remarkable success. Thanks to the availability of vast amounts of image-text pairs on the Internet [5,8,57,61], vision-language pre-training methods have

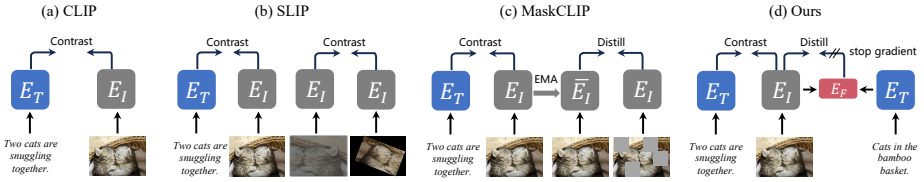

Fig. 1: Pipeline comparison between ours and previous methods. (a) Vanilla CLIP. (b) CLIP + contrastive learning, i.e., SLIP. (c) CLIP + mask self-distillation, i.e., MaskCLIP. (d) Our proposed FuseTeacher. The E_I and E_T are the encoders for image and text respectively. E_F is a *light-weight* fusion encoder. For simplicity, we omit the training loss of E_F in this figure.

been able to learn visual representations that can effectively recognize a wide range of visual concepts. For instance, in the context of zero-shot image classification on the ImageNet *val* set, CLIP [54] demonstrates comparable performance to ImageNet [14] pre-trained ResNet-50 [27], despite CLIP does not utilize the ImageNet *train* set during training.

Vanilla CLIP relies on cross-modality supervision, i.e., learning visual representation from text supervision and vice versa, as shown in Fig. 1 (a). Recent works [16, 37, 42, 47] suggest that this is inadequate as it fails to exploit the full potential of image-text datasets. They further introduce intra-modality supervision to provide more guidance for the visual representation learning, such as SLIP [47] in Fig. 1 (b) and MaskCLIP [16] in Fig. 1 (c). SLIP introduces contrastive learning [9] in image modality, and MaskCLIP combines masked image modelling [25] with CLIP. These methods learn visual representation that is invariant to pre-defined image transformations [33], such as rotation, crop, color jittering, and occlusion.

Nevertheless, both CLIP and its self-supervised variants solely rely on cross-modality and intra-modality supervision. The potential benefits of modality-fused supervision are overlooked. The modality-fused representation augments the image representation with textual information, thus it can be more discriminative and is potential to provide strong supervision. Besides that, CLIP and its variant methods assume a one-to-one correspondence between images and texts. Given a batch of image-text pairs, they pull the paired images and texts close while pushing the unpaired ones apart. However, it is possible that unpaired images and texts in the batch are still related. Pushing them apart introduces noise during training.

In this paper, we conduct experiments to validate that the modality-fused representation is more discriminative than the single-modality representation. Besides that, we propose the FuseTeacher, which effectively learns visual representation through modality-fused supervision. Specifically, the FuseTeacher employs a *lightweight* fusion encoder to enrich the image embedding with text embedding, resulting in a fusion embedding. The fusion embedding can be leveraged to supervise the training of the image encoder through knowledge distilla-

tion. Since classification and retrieval objectives are commonly used for visual representation learning [9, 26, 27, 36, 54, 59], we introduce the following two distillation tasks to transfer the knowledge from the fusion representation to the visual representation: (i) *Classification Distillation*. We use the fusion representation to cluster image-text pairs into K clusters and assign each pair a soft cluster assignment. This cluster assignment acts as a pseudo classification label to supervise the image encoder. (ii) *Retrieval Distillation*. We compute the similarities between the fusion representation and all text representations within the same batch. This similarity distribution serves as a pseudo retrieval similarity, allowing us to avoid pushing away relevant but unpaired images and texts during contrastive learning. The FuseTeacher is easy to corporate with existing vision-language frameworks like CLIP and MaskCLIP. Experiments show that introducing FuseTeacher can bring significant improvements.

Our contribution are three-fold:

- We propose a novel method named FuseTeacher, which introduces modality-fused supervision for visual representation learning with image-text datasets.
- We produce pseudo ground-truth similarities between images and texts by the FuseTeacher, which can be used to prevent the one-to-one contrastive learning from separating unpaired but relevant images and texts.
- We conduct comprehensive experiments to validate the effectiveness of the proposed FuseTeacher. Experiments show that the FuseTeacher can bring consistent and significant improvement over corresponding baseline.

2 Related Work

2.1 Contrastive Vision-language Pre-training

CLIP [54] and its variants [30–32, 39–41, 43, 45, 53, 62, 65, 71–73] have achieved great success recently by utilizing large-scale image-text datasets for pre-training. After pre-training, they achieve state-of-the-art performance on various downstream tasks, including zero-shot image classification, image-text retrieval, and linear probe. However, CLIP learns visual representation with only textual supervision. The text only describes small part of content of corresponding image [16]. Meanwhile, CLIP assumes a one-to-one correspondence between images and texts. It pulls paired image and text close and pushes the unpaired ones away. However, it is possible that some unpaired images and texts are also relevant. Pushing them away will introduce noise during training. In this paper, we propose a novel method named FuseTeacher to tackle these two problems.

2.2 CLIP meets Self-supervised Learning

Some previous methods introduce self-supervised learning in CLIP training [1, 6, 7, 9, 11, 20, 24–26]. SLIP [47] combines SimCLR [9] and CLIP [54] to learn better visual representation. MaskCLIP [16] incorporates Masked Image Modeling [1, 25] into vision language pre-training. The self-supervised learning methods effectively improve the data efficiency. However, these methods only take

cross-modality and intra-modality supervision into consideration, they overlook the potential benefits of modality-fused supervision. Although RA-CLIP [65] enhances image representations through a multi-modal fusion approach, it introduces additional computational overhead during inference. In this paper, we propose the FuseTeacher that introduces modality-fused supervision to guide the visual representation learning. Experimental results show that the proposed FuseTeacher is compatible with CLIP and its self-supervised vairant methods.

2.3 SSL with Self-Distillation

Self-distillation aims at distilling the knowledge from a model itself. In self-supervised learning (SSL), self-distillation is widely used. DINO [7], BYOL [24] and MaskCLIP [16] use momentum encoder as teacher, and train the student model online. Momentum encoder is typical better than the online students since it can be considered as a temporal ensemble of online students [60]. In this paper, we propose a novel teacher model, which takes multi-modal input and produce modality-fused representation. Since the modality-fused representation augments the image representation with text information, it can be more discriminative. We also validate this in our paper by experiments. With the modality-fused model as teacher, the FuseTeacher brings consistent improvements over the CLIP and MaskCLIP baseline.

2.4 Synthetic Image-Text Datasets

Some previous works [3, 18, 38, 42, 44] propose to use large language models or image caption models to generate more captions for exsiting image-text datasets. BLIP [38] proposes CapFilt, which leverages image caption model to produce more text captions for images, and further filters the generated text caption with a filter model to remove noisy image-text pairs. DeCLIP [42] introduces more captions for image by EDA [63] and text retrieval. LaCLIP [18] employs large language model to generate more diverse captions. LLaVA [44] utilizes large language models to generate multi-modal instruct-following data. DALL-E 3 [3] also uses image caption model to obtain more discriptive caption. In this paper, we use LLaVA [44] to generate additional caption for existing image-text. Our processed datasets contains two captions for each image. One is fused with the image embedding and produce the fusion embedding. The other one is used to conduct contrastive learning. Both the baseline and the proposed method use the same datasets for a fair comparison.

3 Method

In this section, we introduce the proposed FuseTeacher. Figure 2 illustrates the training and inference pipeline for the FuseTeacher. During training, the FuseTeacher takes an image and two associated texts as inputs. We randomly select one text for contrastive learning, and use another text together with the

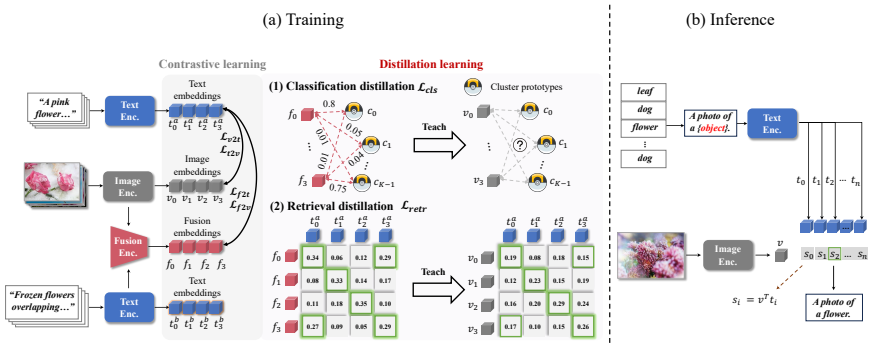

Fig. 2: Overview of the proposed framework. (a) The training stage of FuseTeacher. (b) Once trained, the model only keep the image encoder and the text encoder.

image to produce fusion embedding by a *lightweight* fusion encoder. Then, we utilize the fusion embedding to provide modality-fused supervision for the image encoder with distillation learning. During inference, we can discard the fusion encoder and use the image encoder and text encoder similar to CLIP.

3.1 Image-Text Preprocessing

Large-scale image-text datasets such as YFCC [61], LAION [57] and COYO [5] usually have one text caption for each image. As discussed in Sec. 2.4, generating additional captions for pre-training becomes a common strategy to extend the data annotations, which is adopted in many existing methods such as BLIP [38], LLaVA [44], LaCLIP [18], DeCLIP [42] and DALL-E 3 [3]. In this paper, we employ LLaVA [44] to generate an additional text description for each image in the dataset, and takes two captions and an image as input during training. Since the dataset is extended, it is crucial to mention that all the methods in our experiments are trained on the same dataset for a fair comparison.

3.2 Feature Extraction

Given N images $\{\mathbf{I}_i\}_{i=1}^N$and $2N$ corresponding texts $\{\mathbf{T}_i^a, \mathbf{T}_i^b\}_{i=1}^N$ as input, the FuseTeacher first extract feature embeddings for both images and texts. The images are fed into the image encoder $\mathcal{E}_I$ to produce image embeddings $\{\mathbf{v}_i\}_{i=1}^N$:

$$\mathbf{v}_i = \mathcal{E}_I(\mathbf{I}_i), \tag{1}$$

where $\mathbf{v}_i \in \mathbb{R}^d$ and d is the dimension of the embeddings. The texts $\{\mathbf{T}_i^a, \mathbf{T}_i^b\}_{i=1}^N$ are fed into the text encoder $\mathcal{E}_T$ to obtain text embeddings $\{\mathbf{t}_i^a\}_{i=1}^N$ and $\{\mathbf{t}_i^b\}_{i=1}^N$:

$$\begin{aligned} \mathbf{t}_i^a &= \mathcal{E}_T(\mathbf{T}_i^a), \\ \mathbf{t}_i^b &= \mathcal{E}_T(\mathbf{T}_i^b), \end{aligned} \tag{2}$$

where $\mathbf{t}_i^a \in \mathbb{R}^d$ and $\mathbf{t}_i^b \in \mathbb{R}^d$.

Besides that, given a pair of image and a randomly selected text, the FuseTeacher feeds the embeddings of all image patches and text tokens into the fusion encoder, and obtain a fusion embedding $\mathbf{f}_i \in \mathbb{R}^d$. Following previous works [38,39], the fusion encoder is constructed by few transformer blocks with cross-attention to interact the image patches and text token embeddings. More details of the fusion encoder can be found in the appendix. The fusion embedding $\mathbf{f}_i$ is expected to be more powerful than $\mathbf{v}_i$ and $\mathbf{t}_i$, since it contains multi-modality information and could explore fine-grained cross-modal interaction between the image patches and text tokens. The experiments in Sec. 4.2 show that using the fusion embedding could achieve significant performance gains for classification and retrieval, and similar conclusions are observed in previous methods [39,67]. Thus we use $\mathbf{f}_i$ to supervise the learning of $\mathcal{E}_I$, which will be discussed later.

3.3 Training

The training objective of FuseTeacher contains two parts:

- Contrastive Learning: Training the image encoder, text encoder and fusion encoder with contrastive learning.
- Distillation Learning: Distilling the knowledge of fusion representation to visual representation.

During training, contrastive learning and distillation learning are applied simultaneously to train the FuseTeacher model end-to-end.

Contrastive Learning The loss function used in contrastive learning is:

$$\mathcal{L}_{ct} = \mathcal{L}_{v2t} + \mathcal{L}_{t2v} + \mathcal{L}_{f2t} + \mathcal{L}_{t2f}. \tag{3}$$

$\mathcal{L}_{v2t}$ and $\mathcal{L}_{t2v}$ indicate the contrastive learning between image and text:

$$\mathcal{L}_{v2t} = -log(\frac{exp(\sigma(\mathbf{v}_i, \mathbf{t}_i)/\tau_1)}{\sum_{j=1}^{N} exp(\sigma(\mathbf{v}_i, \mathbf{t}_j)/\tau_1)}), \tag{4}$$

$$\mathcal{L}_{t2v} = -log(\frac{exp(\sigma(\mathbf{t}_i, \mathbf{v}_i)/\tau_1)}{\sum_{j=1}^{N} exp(\sigma(\mathbf{t}_i, \mathbf{v}_j)/\tau_1)}), \tag{5}$$

where $\sigma(\cdot, \cdot)$ calculates the cosine similarity between vectors. τ_1 is a trainable temperature parameter. Here we use $\mathbf{t}_i$ for simplicity since $\mathbf{t}_i^a$ is selected for contrastive learning. $\mathcal{L}_{f2t}$ and $\mathcal{L}_{t2f}$ indicate the contrastive learning between fusion representation and text representation:

$$\mathcal{L}_{f2t} = -log(\frac{exp(\sigma(\mathbf{f}_i, \mathbf{t}_i)/\tau_2)}{\sum_{j=1}^{N} exp(\sigma(\mathbf{f}_i, \mathbf{t}_j)/\tau_2)}), \tag{6}$$

$$\mathcal{L}_{t2f} = -log(\frac{exp(\sigma(\mathbf{t}_i, \mathbf{f}_i)/\tau_2)}{\sum_{j=1}^{N} exp(\sigma(\mathbf{t}_i, \mathbf{f}_j)/\tau_2)}), \tag{7}$$

where τ_2 is another trainable temperature parameter. With Eq. 7 and Eq. 6, the fusion encoder can learn a fusion representation of image and text.

Distillation Learning Besides the contrastive learning, we introduce distillation learning to distill the knowledge of the fusion representation to the visual representation. Traditional visual representation learning methods typically use classification label supervision [27, 36, 59] or contrastive retrieval objectives [9, 26, 54]. Inspired by them, we introduce classification distillation and retrieval distillation to supervise the visual representation learning.

Classification Distillation. As there is no classification labels for the image-text datasets, we produce pseudo labels through clustering. We first construct K cluster prototypes $\mathbf{c}$ and calculate a soft cluster assignment for each image-text pair with the fusion embedding. Given the image of an image-text pair, the image encoder learns the soft-assignment with the following objective:

$$\mathcal{L}_{cls} = -\sum_{i}^{N}\sum_{k}^{K}\mathbf{P}_{i,k}^{f2c} log(\mathbf{P}_{i,k}^{v2c}), \tag{8}$$

where $\mathbf{P}_{i,k}^{f2c}$ is the probability that the i-th fusion embeddings is assigned to the k-th clustering centroids, i.e.,

$$\mathbf{P}_{i,k}^{f2c} = \frac{exp(\sigma(\mathbf{f}_i, \mathbf{c}_k)/\tau_{f2c})}{\sum_{j=1}^{K} exp(\sigma(\mathbf{f}_i, \mathbf{c}_j)/\tau_{f2c})}, \tag{9}$$

where $\mathbf{c}_i$ is a learnable embedding for the i-th clustering centroid, inspired by DINO [7]. Similar formulation holds for $\mathbf{P}_{i,k}^{v2c}$. To avoid trivial solutions where all $\mathbf{f}_i$ collapse into a single clustering centroid or are uniformly assigned to all clustering centroids, we utilize the Sinkhorn-Knopp algorithm to adjust the assignment $\mathbf{P}_{i,k}^{f2c}$ following related works [6,49]. By minimizing $\mathcal{L}_{cls}$ during training, we learn the image encoder to produce embeddings that share similar semantic information with the fusion embeddings. Note that we stop the gradient of $\mathbf{P}_{i,k}^{f2c}$ since it acts as target and doesn't require updates.

Retrieval Distillation. The contrastive loss of CLIP can introduce noise if there are relevant but unpaired image and text in the same batch. To alleviate this problem, we introduce the retrieval distillation objective. Specifically, we use the similarity distribution between fusion embeddings and text embeddings as pseudo label for image-text contrastive learning, and learn the image encoder to approximate this distribution:

$$\mathcal{L}_{retr} = -\sum_{i}^{N}\sum_{j}^{N}\mathbf{P}_{i,j}^{f2t} log(\mathbf{P}_{i,j}^{v2t}) - \sum_{i}^{N}\sum_{j}^{N}\mathbf{P}_{i,j}^{t2f} log(\mathbf{P}_{i,j}^{t2v}), \tag{10}$$

where $\mathbf{P}_{i,j}^{f2t}$ is the similarity between the i-th fusion embedding and the j-th text embedding, i.e.,

$$\mathbf{P}_{i,j}^{f2t} = \frac{exp(\sigma(\mathbf{f}_i, \mathbf{t}_j)/\tau_2)}{\sum_{j=1}^{N} exp(\sigma(\mathbf{f}_i, \mathbf{t}_j)/\tau_2)}. \tag{11}$$

Similar to $\mathbf{P}_{i,j}^{v2t}$, $\mathbf{P}_{i,j}^{t2f}$ and $\mathbf{P}_{i,j}^{t2v}$. Note that we also stop the gradient of $\mathbf{P}_{i,j}^{f2t}$ and $\mathbf{P}_{i,j}^{t2f}$ during training.

Finally, the total loss function of the FuseTeacher is:

$$\mathcal{L} = \mathcal{L}_{ct} + \lambda_1 \mathcal{L}_{cls} + \lambda_2 \mathcal{L}_{retr}. \tag{12}$$

We set λ_1 and λ_2 to 1 to avoid hyper-parameter tuning.

3.4 Inference

After training, we can discard the fusion encoder and only keep the image encoder and the text encoder. This makes us share the same dual encoder pipeline and computation cost with CLIP, as described in Figure 2 (b).

Note that our framework produce a by-product: a multi-modal encoder, which can be used to encode image-text content. Given an image-text pair, the FuseTeacher first extract patch and token embeddings of image and text respectively, then feed them into the fusion encoder to obtain the fusion embedding. Different from CLIP that it can only encodes the image and text separately and subsequently combines the two embeddings through summation, our multi-modal encoder provides better fusion embedding for image-text input. Experimental results can be found in Section 4.2.

4 Experiments

In this section, we first describe the implementation details of our experiments, including the model architecture, datasets and training details. Next, we perform ablation experiments to confirm our hypothesis that the fusion representation is more discriminative than single-modality representation. Furthermore, we conduct experiments to validate the effectiveness of the FuseTeacher, and compare it with previous state-of-the-arts.

4.1 Implementation Details

Architecture. The image encoder is implemented with vision transformer [17]. We use three different vision transformers in our experiments, i.e., ViT-B/32, ViT-B/16, and ViT-L/14, following previous works [54]. The input image is of size 224×224, and is split to 32×32 (ViT-B/32), 16×16 (ViT-B/16), or 14×14 (ViT-L/14) patches. An additional learnable [CLS] token embedding is concatenated to the patch embeddings before being fed into the image encoder. The text encoder is a BERT-like encoder [15], which also comprises 12 layers, with a width of 768 and 12 heads. The number of text tokens is limited to 77 with necessary truncations or paddings. The fusion encoder consists of 2 layer of transformer blocks, with the same width and heads of image encoder. K in Eq. 8 is set to 4096 in all experiments. Ablation experiments on different numbers of fusion layers and K can be found in the appendix.

Datasets. Following recent works, we utilize YFCC15M [61] for most of our experiments, which consists of 15 million of image-text pairs. Inspired by related

works [3, 18, 38, 42, 44], we employ LLaVA [44] to generate an extra caption for each image. Consequently, this dataset has two texts per image, we refer to it as YFCC15M-Cap in this paper.

There exists some image-text datasets with hundreds of million image-text pairs, e.g., COYO [5], and LAION 2B [56]. Training models with these datasets is not affordable for most laboratories and companies. To address this issue, we follow recent works [22,49,66] and sample two subsets from them. Specifically, we create two new datasets by sampling from existing collections: Union23M with 23 million image-text pairs from LAION 400M [57], COYO [5], YFCC15M [61], CC12M [8], CC3M [58], and SBU [50], and Union65M with 65 million pairs from LAION 2B [56], COYO [5], YFCC15M [61], CC12M [8], CC3M [58], and SBU [50]. The detail can be found in the appendix.

For the evaluation datasets, we use six ImageNet series datasets, i.e., ImageNet-1K [14], ImageNet V2 [55], ImageNet-R [28], ImageNet-A [29], ImageNet-S [23] and ObjectNet [2]. Besides that, some other commonly used datasets are also used, including Microsoft COCO [10], Flickr30K [70], CIFAR 10 [35], CIFAR 100 [35], Oxford Pets [51], SUN 397 [64], Food 101 [4], Flowers [48], Standford Cars [34], Caltech 101 [21], Aircraft [46] and DTD [12].

Training. We implement our framework with PyTorch [52]. For the experiments on YFCC15M-Cap, we train all models for 25 epochs if not specified, following MaskCLIP [16]. For the experiments on Union23M and Union65M, we train all models for 32 epochs, following OpenAI CLIP [54]. The learning rate is initialized to 2.5e-3, with a linear warm up strategy during the first 5 epochs, and decay to 0 with a cosine schedule. LAMB [69] optimizer is used with the betas equal to [0.9, 0.98] and weight decay equal to 0.2. The batch size is set to 4096 following previous methods [16, 37, 42, 47, 68]. Strong data augmentation is used during pre-training, including random crop, random color distortions, random gaussian blur and RandAugment [13], following previous works [9, 42].

For the linear-probe experiments, we train all models with SGD optimizer for 90 epochs. The momentum is set to 0.9. The learning rate is initialized to 0.1 and decays to 0 with a cosine learning rate scheduler. The weight decay is set to 0. For the k-NN classification implementation, we use the code of DINO [7]. More details of downstream evaluation, training efficiency are in the appendix.

4.2 Fusion Embedding is More Discriminative

Intuitively, modality-fused encoders are expected to outperform single-modality encoders due to their ability to encode more comprehensive information. In this section, we present experiments designed to substantiate this hypothesis. Specifically, we conduct a multi-modal zero-shot classification task and two multi-modal retrieval tasks: image+text to text retrieval (M2T) and text to image+text retrieval (T2M). For the multi-modal zero-shot classification task, we utilize image+text as input, and conduct zero-shot classification as usual. For the image+text to text retrieval task, we utilize image+text pairs as queries and search for texts that are relevant to the given queries. Conversely, for the text to im-

Table 1: Results of multi-modal zero-shot classification on ImageNet-MM datasets (%), and multi-modal retrieval on Flickr-MM and COCO-MM datasets (%).

	modality-fused	ImageNet-MM		Flickr30K-MM		COCO-MM	
		top-1	top-5	M2T	T2M	M2T	T2M
CLIP Image	✗	65.2	87.3	68.4	53.0	49.7	35.6
CLIP Text	✗	67.7	77.2	75.3	48.2	39.0	28.1
CLIP Image+Text	✓	74.2	86.5	84.1	62.1	49.8	41.1
FuseTeacher	✓	**78.3**	**94.9**	**86.7**	**72.6**	**62.3**	**46.7**

age+text retrieval task, we use text inputs as queries and search for image+text pairs that correspond to the queries.

To facilitate these experiments, we introduce three datasets based on the ImageNet-Caption [19], Flickr30K [70] and MS-COCO [10] datasets. Since most ImageNet images are collected from Flickr, the ImageNet-Caption dataset obtains the original text caption for the ImageNet images with the Flickr API. We sample 50 images for each category, and construct a new dataset referred as ImageNet-MM. This dataset will be used for the multi-modal zero-shot classification evaluation. Besides that, we construct two multi-modal retrieval datasets with Flickr30K and MS-COCO. Each of these two datasets contains five captions for each image. From these captions, we randomly select one and combine it with the corresponding image to create the image+text inputs. The remaining four captions are used as individual text inputs. This process allows us to construct two new multi-modal retrieval datasets, namely Flickr-MM and COCO-MM.

We use the ViT-B/32 based CLIP and FuseTeacher trained on the Union23M dataset for evaluation. Experimental results are presented in Table 1. The *CLIP Image* model encodes only the image component of the multi-modal input, while the *CLIP Text* model encodes only the text component. The *CLIP Image+Text* model encodes the image and text components separately. The resulting embeddings are then summed, normalized, and used as multi-modal embeddings. On the other hand, the *FuseTeacher* model encodes the image+text input using the fusion encoder. The experimental results demonstrate that *CLIP Image+Text* and *FuseTeacher* significantly outperform *CLIP Image* and *CLIP text*, providing empirical evidence that modality-fused encoders are indeed superior to single-modality encoders. Furthermore, the FuseTeacher model achieves the best performance, thereby validating the effectiveness of the fusion encoder.

4.3 Comparison with CLIP and MaskCLIP Baseline

To validate the effectiveness of the FuseTeacher, we incorporate it with two typical frameworks, i.e., CLIP [54] and MaskCLIP [16]. CLIP is the most widely used framework in contrastive language-image pre-training. MaskCLIP is a recent state-of-the-art method that combines CLIP with self-supervised learning.

Experiments on CLIP baseline. We first compare the FuseTeacher with CLIP baseline. Both methods are pre-trained on the YFCC15M-Cap dataset. For

Table 2: Experiments on the YFCC15M-Cap dataset. FuseTeacher+ denotes applying FuseTeacher on MaskCLIP.

	IN	IN V2	IN-R	IN-A	IN-S	ObjectNet	Flickr30K		COCO		**Avg.**
							I2T	T2I	I2T	T2I	
CLIP [54]	41.1	35.7	36.7	12.2	16.2	18.7	74.9	51.6	47.3	29.4	36.4
FuseTeacher	**43.5**	**38.1**	**40.4**	**13.9**	**17.7**	**20.6**	**76.2**	**54.5**	**48.4**	**31.1**	**38.4**
MaskCLIP [16]	43.1	37.2	38.4	14.1	16.9	21.3	77.1	54.5	49.4	31.3	38.3
FuseTeacher+	**44.9**	**39.0**	**42.5**	**16.4**	**19.4**	**22.3**	**78.1**	**56.7**	**50.5**	**32.7**	**40.3**

the CLIP pre-training, we random sample a caption from the two available captions for a given image in each epoch, and apply the aforementioned strong data augmentation for the image. The experimental results can be found in Table 2. The FuseTeacher brings a 2.0% average improvement over the CLIP baseline across six ImageNet variant datasets and two cross-modal retrieval datasets. We attribute the improvement to the following two aspects: (i) The classification distillation provides reliable soft-assignment between the input image and different concepts (clustering centroids), introducing more informative supervision. (ii) The retrieval distillation establishes effective similarities between all images and texts using a strong fusion representation. This prevents the model from pushing away relevant but unpaired images and texts.

Experiments on MaskCLIP baseline. To address the limitation that text captions only provide partial information about the corresponding image, previous methods such as MaskCLIP have proposed incorporating intra-modality supervision in contrastive language-image pre-training. Our proposed FuseTeacher is compatible with these methods. To validate that, we integrate the FuseTeacher with MaskCLIP, referred to as FuseTeacher+. The results are presented in Table 2. We observe a further performance improvement of 2.0% when compared to the strong MaskCLIP baseline.

4.4 Comparison with More CLIP Vairant Methods

We also compare our framework with more recent CLIP variant methods, including MS-CLIP [68], SLIP [47], DeCLIP [42], and UniCLIP [37]. Following UniCLIP [37], the comparison is conducted on 13 commonly used datasets. These datasets consist of 11 zero-shot image classification and linear probe benchmarks, as well as 2 cross-modal retrieval benchmarks. The experimental results can be found in Table 3 and Table 4. The FuseTeacher demonstrates superior performance on all 13 datasets.

4.5 Ablation on Distillation Loss

The FuseTeacher incorporates two distillation losses: classification distillation loss and retrieval distillation loss. To assess the effectiveness of each distillation loss, we conduct experiments with and without these two losses. The evaluation is performed on the ImageNet-1K *val* set. In addition to reporting the zero-shot

Table 3: Zero-shot image classification performance and linear probe classification performance on 11 downstream datasets (%).

	Method	Pretrain	Epochs	ImageNet	Pets	CIFAR10	CIFAR100	SUN397	Food101	Flowers	Car	Caltech101	Aircraft	DTD	**Average**
zero-shot	CLIP [54]	YFCC15M	50	31.3	19.4	62.3	33.6	40.2	33.7	6.3	2.1	55.4	1.4	16.9	27.5
	MS-CLIP [68]	YFCC15M	50	36.7	–	–	–	–	–	–	–	–	–	–	–
	SLIP [47]	YFCC15M	50	38.3	28.3	72.2	45.3	45.1	44.7	6.8	2.9	65.9	1.9	21.8	33.9
	DeCLIP [42]	YFCC15M	50	41.2	30.2	72.1	39.7	51.6	46.9	7.1	3.9	70.1	2.5	24.2	35.4
	UniCLIP [37]	YFCC15M	50	42.8	32.5	78.6	47.2	50.4	48.7	8.1	3.4	73.0	2.8	23.3	37.3
	CLIP [54]	YFCC15M-Cap	25	41.1	38.5	77.9	54.9	53.5	51.6	53.4	8.4	73.4	9.3	33.2	45.0
	MaskCLIP [16]	YFCC15M-Cap	25	43.1	39.6	78.9	53.9	54.6	53.6	55.4	9.3	74.5	9.3	37.1	46.3
	FuseTeacher	YFCC15M-Cap	25	43.5	40.7	81.1	57.3	55.6	52.9	53.3	9.5	76.3	9.3	35.3	46.8
	FuseTeacher+	YFCC15M-Cap	25	44.9	39.9	83.3	58.3	56.1	57.1	56.0	9.5	77.9	10.0	36.1	48.1
	FuseTeacher+	YFCC15M-Cap	50	**48.5**	**47.5**	**87.9**	**65.0**	**58.5**	**62.5**	**59.2**	**11.6**	**80.3**	**10.7**	**38.7**	**51.9**
linear probe	CLIP [54]	YFCC15M	50	61.1	71.2	89.2	72.1	70.1	71.4	93.2	34.9	84.3	29.7	60.9	67.1
	MS-CLIP [68]	YFCC15M	50	63.7	62.1	87.2	66.7	71.7	76.0	93.8	27.5	81.6	32.9	69.4	66.6
	SLIP [47]	YFCC15M	50	68.1	75.4	90.5	75.3	73.5	77.1	96.1	43.0	87.2	34.1	71.1	71.9
	DeCLIP [42]	YFCC15M	50	69.2	76.5	88.6	71.6	75.9	79.3	96.7	42.6	88.0	32.6	69.1	71.8
	UniCLIP [37]	YFCC15M	50	70.8	**83.1**	92.5	78.2	**77.0**	**81.3**	**97.1**	49.8	88.9	36.2	72.8	75.2
	CLIP [54]	YFCC15M-Cap	25	66.5	77.3	93.2	77.0	68.3	73.7	95.4	51.9	89.2	51.4	73.4	74.3
	MaskCLIP [16]	YFCC15M-Cap	25	68.2	79.9	93.1	77.1	69.3	75.8	95.7	52.0	90.3	51.9	73.6	75.2
	FuseTeacher	YFCC15M-Cap	25	68.4	79.5	94.3	78.3	69.6	75.5	95.7	54.0	90.0	53.4	74.5	75.7
	FuseTeacher+	YFCC15M-Cap	25	69.7	79.9	94.0	78.6	70.5	77.3	95.9	51.4	90.9	52.8	73.9	75.9
	FuseTeacher+	YFCC15M-Cap	50	**72.4**	81.4	**95.7**	**81.3**	73.6	80.7	96.9	**55.0**	**91.5**	**53.4**	**76.8**	**78.1**

Table 4: Results of zero-shot image-text retrieval on Flickr30K and MS-COCO (%).

	Pretrain	Flickr30K						MS-COCO					
		Image-to-text			Text-to-image			Image-to-text			Text-to-image		
		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 [54]	YFCC15M	37.3	66.2	77.1	24.9	49.0	60.0	20.1	42.9	55.1	13.3	31.7	42.3
SLIP [47]	YFCC15M	48.7	75.2	84.7	33.1	59.0	68.8	26.9	51.9	63.8	18.2	39.6	51.1
DeCLIP [42]	YFCC15M	51.3	79.3	88.7	34.8	62.2	71.5	28.1	53.6	65.2	17.9	39.8	51.6
UniCLIP [37]	YFCC15M	55.7	82.9	90.0	36.7	62.6	72.4	32.0	58.8	70.3	20.3	43.1	54.5
CLIP [54]	YFCC15M-Cap	74.9	92.4	96.8	51.6	78.2	85.7	47.3	73.3	82.4	29.4	55.1	66.4
MaskCLIP [16]	YFCC15M-Cap	77.1	93.1	**97.3**	54.5	80.0	87.1	49.4	76.2	84.7	31.3	57.3	68.6
FuseTeacher	YFCC15M-Cap	76.2	92.2	95.8	54.5	80.0	87.1	48.4	74.0	82.8	31.1	57.2	68.4
FuseTeacher+	YFCC15M-Cap	**78.1**	**94.3**	97.1	**56.7**	**82.8**	**88.9**	**50.5**	**77.1**	**85.0**	**32.7**	**58.6**	**69.7**

classification and linear-probe accuracy, we also present the k-NN classification accuracy, following recent self-supervised learning methods [7, 49, 74].

The results are summarized in Table 5. From the results, we can observe that both distillation losses contribute to the overall improvements. The retrieval distillation performs slightly better on the zero-shot classification task since it directly supervises the relation between images and texts. However, the classification distillation, which learns to group visual embeddings for the same concept, performs better than the retrieval distillation loss on the linear-probe and k-NN classification tasks. Combining both distillation losses leads to the best overall performance.

4.6 Experiments on Large backbones and More Data

Pre-training on larger datasets like LAION400M, COYO, and LAION2B typically yields improvements. However, training on these dataset needs lots of GPUs, which is not affordable for most laboratories and companies. Recent

Table 5: Ablation on distillation losses (%).

classification distill	retrieval distill	zero-shot	linear-probe	k-NN
✗	✗	41.1	66.5	59.1
✓	✗	42.9	67.6	60.9
✗	✓	43.0	67.4	60.5
✓	✓	**43.5**	**68.4**	**61.5**

Table 6: Experiments on larger backbone and more data.

Method	Backbone	Dataset	IN	IN v2	IN-R	IN-A	IN-S	ObjectNet	Flickr30K		COCO		**Avg.**
									I2T	T2I	I2T	T2I	
CLIP [54]	ViT-B/32	Union23M	65.3	56.2	65.7	17.4	45.7	42.7	73.8	55.0	51.4	35.3	50.9
FuseTeacher	ViT-B/32	Union23M	67.1	58.2	72.4	21.4	48.9	46.9	74.8	56.8	53.3	37.4	53.7
CLIP [54]	ViT-B/16	Union23M	68.7	60.5	70.1	29.3	47.9	52.0	77.3	61.4	52.7	39.4	55.9
FuseTeacher	ViT-B/16	Union23M	71.4	63.1	79.5	40.2	53.7	58.5	79.5	63.7	55.5	40.6	60.6
CLIP [54]	ViT-L/14	Union23M	72.5	64.9	77.7	43.1	54.2	58.9	82.6	65.5	56.0	42.0	61.7
MaskCLIP [16]	ViT-L/14	Union23M	74.2	67.5	83.4	56.2	57.3	65.7	84.1	69.3	56.9	44.1	65.9
FuseTeacher	ViT-L/14	Union23M	73.6	65.9	82.3	50.8	57.3	62.0	84.3	67.8	56.3	42.9	64.3
FuseTeacher+	ViT-L/14	Union23M	74.9	67.7	85.2	57.9	59.3	65.7	86.3	70.1	58.6	44.4	67.0
FuseTeacher+	ViT-L/14	Union65M	77.1	70.1	87.7	62.1	62.8	69.4	87.8	74.5	60.1	46.9	69.9

works such as CiT [66], DataComp [22], and DINO v2 [49] propose sampling images or image-text pairs to tackle this problem. We follow them and sample two new datasets named Union23M and Union65M, as detailed in Section 4.1. We first compare the FuseTeacher with CLIP by pre-training on the Union23M dataset with different scales of vision backbones. Experimental results show that CLIP pre-trained on the Union23M dataset performs much better than that pre-trained on YFCC15M-Cap, establishing a stronger baseline. Nevertheless, the proposed FuseTeacher still outperforms the CLIP baseline across different scale of vision backbones. Additionally, we train MaskCLIP on the Union23M dataset. Due to GPU limitations, we only train one MaskCLIP model with ViT-L/14 as the vision backbone. Experimental result show that FuseTeacher+ achieves better performance than MaskCLIP.

After comparing FuseTeacher+ with CLIP and MaskCLIP on the Union23M dataset with different vision backbone, we also scale up the training data and pre-train FuseTeacher+ on the Union65M dataset. By pre-training on more data, the FuseTeacher+ further obtains a 2.9% average improvement, showing that the model can benefit from larger datasets. Our pre-trained models will be released.

4.7 Visualization of the Retrieval and Classification Pseudo Labels

CLIP and its variant methods adopt contrastive loss for images and texts in the same batch. The contrastive loss pull paired images and texts close and push the unpaired ones away. CLIP typically uses a very large batch size during training, such as 4096 or even larger. However, it is possible that unpaired images and text are also relevant. Pushing these false negative pairs will introduce noise during training. Is this section, we provide some visualization of the false negative pairs. Specifically, we save the text, its paired image, and its relevant but unpaired

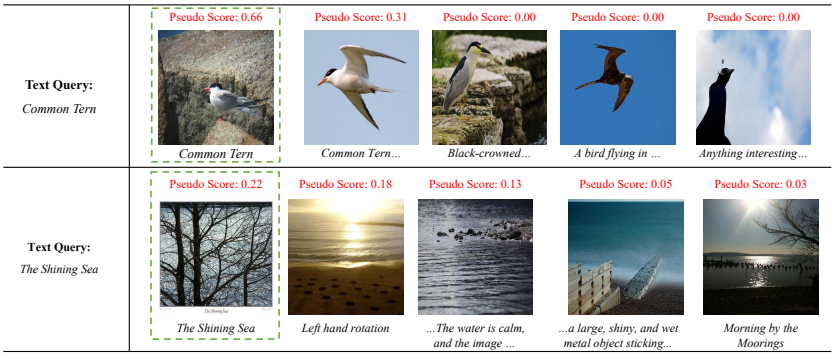

Fig. 3: Visualization of some relevant but unpaired image-text pairs in the same batch. For each text query, we demonstrate its top-5 most similar image-text pairs in the same batch. The image in the dashed green box is the paired image of the text query.

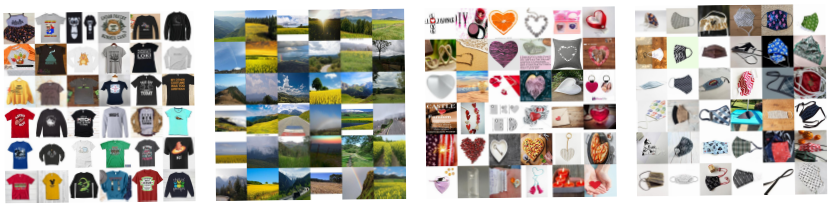

Fig. 4: Clustering results of the Union23M dataset.

images during training. The results are shown in Fig. 3. We can see that there actually exists some relevant images that are unpaired with the corresponding text. We also present the pseudo label $\mathbf{P}_{i,j}^{t2f}$ used in Eq. 10, the pseudo labels yield more accurate supervision than traditional contrastive learning approaches.

We also visualize the pseudo classification labels. Fig. 4 shows that the images of the same pseudo classification labels share similar appearance and semantic information. Incorporating this classification supervision during training helps the image encoder learn more discriminative representations.

5 Conclusion

The FuseTeacher is proposed in this paper. Different from previous methods that only exploit cross-modality and intra-modality supervision, the FuseTeacher introduces modality-fused supervision for better visual representation learning. The modality-fused supervision is utilized by two distillation ways: classification distillation and retrieval distillation. The FuseTeacher is compatible with existing contrastive language-image pre-training methods such as CLIP and MaskCLIP. Experimental results on various tasks validate the effectiveness of the modality-fused supervision, including zero-shot classification, linear-probe, k-NN, and cross-modal retrieval.

References

1. Bao, H., Dong, L., Piao, S., Wei, F.: BEiT: BERT pre-training of image Transformers (2021)
2. Barbu, A., Mayo, D., Alverio, J., Luo, W., Wang, C., Gutfreund, D., Tenenbaum, J., Katz, B.: ObjectNet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. In: Adv. Neural Inform. Process. Syst. pp. 9448–9458 (2019)
3. Betker, J., Goh, G., Jing, L., Brooks, T., Wang, J., Li, L., Ouyang, L., Zhuang, J., Lee, J., Guo, Y., Manassra, W., Dhariwal, P., Chu, C., Jiao, Y., Ramesh, A.: Improving image generation with better captions (2023), `https://openai.com/dall-e-3`
4. Bossard, L., Guillaumin, M., Gool, L.V.: Food-101–mining discriminative components with random forests. In: Eur. Conf. Comput. Vis. pp. 446–461 (2014)
5. Byeon, M., Park, B., Kim, H., Lee, S., Baek, W., Kim, S.: COYO-700M: Image-text pair dataset. `https://github.com/kakaobrain/coyo-dataset` (2022)
6. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. In: Adv. Neural Inform. Process. Syst. pp. 9912–9924 (2020)
7. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision Transformers. In: Int. Conf. Comput. Vis. pp. 9650–9660 (2021)
8. Changpinyo, S., Sharma, P., Ding, N., Soricut, R.: Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 3558–3568 (2021)
9. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: Int. Conf. Mach. Learn. pp. 1597–1607 (2020)
10. Chen, X., Fang, H., Lin, T.Y., Vedantam, R., Gupta, S., Dollár, P., Zitnick, C.L.: Microsoft COCO captions: Data collection and evaluation server (2015)
11. Chen, X., He, K.: Exploring simple siamese representation learning. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 15750–15758 (2021)
12. Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., Vedaldi, A.: Describing textures in the wild. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 3606–3613 (2014)
13. Cubuk, E.D., Zoph, B., Shlens, J., Le, Q.V.: RandAugment: Practical automated data augmentation with a reduced search space. In: IEEE Conf. Comput. Vis. Pattern Recog. Worksh. pp. 702–703 (2020)
14. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 248–255 (2009)
15. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional Transformers for language understanding. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 4171–4186 (2019)
16. Dong, X., Bao, J., Zheng, Y., Zhang, T., Chen, D., Yang, H., Zeng, M., Zhang, W., Yuan, L., Chen, D., et al.: MaskCLIP: Masked self-distillation advances contrastive language-image pretraining. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 10995–11005 (2023)
17. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth

16x16 words: Transformers for image recognition at scale. In: Int. Conf. Learn. Represent. (2020)
18. Fan, L., Krishnan, D., Isola, P., Katabi, D., Tian, Y.: Improving CLIP training with language rewrites. In: Adv. Neural Inform. Process. Syst. (2023)
19. Fang, A., Ilharco, G., Wortsman, M., Wan, Y., Shankar, V., Dave, A., Schmidt, L.: Data determines distributional robustness in contrastive language image pre-training (clip). In: Int. Conf. Mach. Learn. pp. 6216–6234 (2022)
20. Fang, Y., Wang, W., Xie, B., Sun, Q., Wu, L., Wang, X., Huang, T., Wang, X., Cao, Y.: EVA: exploring the limits of masked visual representation learning at scale. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 19358–19369 (2023)
21. Fei-Fei, L., Fergus, R., Perona, P.: Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In: IEEE Conf. Comput. Vis. Pattern Recog. Worksh. pp. 178–178 (2004)
22. Gadre, S.Y., Ilharco, G., Fang, A., Hayase, J., Smyrnis, G., Nguyen, T., Marten, R., Wortsman, M., Ghosh, D., Zhang, J., Orgad, E., Entezari, R., Daras, G., Pratt, S.M., Ramanujan, V., Bitton, Y., Marathe, K., Mussmann, S., Vencu, R., Cherti, M., Krishna, R., Koh, P.W., Saukh, O., Ratner, A., Song, S., Hajishirzi, H., Farhadi, A., Beaumont, R., Oh, S., Dimakis, A., Jitsev, J., Carmon, Y., Shankar, V., Schmidt, L.: DataComp: In search of the next generation of multimodal datasets (2023)
23. Gao, S., Li, Z., Yang, M., Cheng, M., Han, J., Torr, P.H.S.: Large-scale unsupervised semantic segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **45**(6), 7457–7476 (2023)
24. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al.: Bootstrap your own latent-a new approach to self-supervised learning. In: Adv. Neural Inform. Process. Syst. pp. 21271–21284 (2020)
25. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 16000–16009 (2022)
26. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 9729–9738 (2020)
27. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 770–778 (2016)
28. Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., Song, D., Steinhardt, J., Gilmer, J.: The many faces of robustness: A critical analysis of out-of-distribution generalization. In: Int. Conf. Comput. Vis. pp. 8320–8329 (2021)
29. Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., Song, D.: Natural adversarial examples. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 15262–15271 (2021)
30. Ilharco, G., Wortsman, M., Wightman, R., Gordon, C., Carlini, N., Taori, R., Dave, A., Shankar, V., Namkoong, H., Miller, J., Hajishirzi, H., Farhadi, A., Schmidt, L.: Openclip (2021). `https://doi.org/10.5281/zenodo.5143773`, `https://doi.org/10.5281/zenodo.5143773`
31. Jeong, J., Zou, Y., Kim, T., Zhang, D., Ravichandran, A., Dabeer, O.: WinCLIP: Zero-/few-shot anomaly classification and segmentation. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 19606–19616 (June 2023)
32. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning

with noisy text supervision. In: Int. Conf. Mach. Learn. pp. 4904–4916. PMLR (2021)
33. Kong, X., Zhang, X.: Understanding masked image modeling via learning occlusion invariant feature. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 6241–6251 (2023)
34. Krause, J., Stark, M., Deng, J., Fei-Fei, L.: 3d object representations for fine-grained categorization. In: Int. Conf. Comput. Vis. Worksh. pp. 554–561 (2013)
35. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009)
36. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: Adv. Neural Inform. Process. Syst. (2012)
37. Lee, J., Kim, J., Shon, H., Kim, B., Kim, S.H., Lee, H., Kim, J.: UniCLIP: Unified framework for contrastive language-image pre-training. In: Adv. Neural Inform. Process. Syst. pp. 1008–1019 (2022)
38. Li, J., Li, D., Xiong, C., Hoi, S.C.H.: BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In: Int. Conf. Mach. Learn. pp. 12888–12900 (2022)
39. Li, J., Selvaraju, R.R., Gotmare, A.D., Joty, S.R., Xiong, C., Hoi, S.C.H.: Align before fuse: Vision and language representation learning with momentum distillation. In: Adv. Neural Inform. Process. Syst. (2021)
40. Li, P., Xie, C.W., Xie, H., Zhao, L., Zhang, L., Zheng, Y., Zhao, D., Zhang, Y.: MomentDiff: Generative video moment retrieval from random to real. In: Adv. Neural Inform. Process. Syst. pp. 65948–65966 (2023)
41. Li, P., Xie, C.W., Zhao, L., Xie, H., Ge, J., Zheng, Y., Zhao, D., Zhang, Y.: Progressive spatio-temporal prototype matching for text-video retrieval. In: Int. Conf. Comput. Vis. pp. 4100–4110 (2023)
42. Li, Y., Liang, F., Zhao, L., Cui, Y., Ouyang, W., Shao, J., Yu, F., Yan, J.: Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. In: Int. Conf. Learn. Represent. (2022)
43. Liu, F., Chen, D., Guan, Z., Zhou, X., Zhu, J., Zhou, J.: RemoteCLIP: A vision language foundation model for remote sensing (2023)
44. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning (2023)
45. Luo, H., Ji, L., Zhong, M., Chen, Y., Lei, W., Duan, N., Li, T.: CLIP4Clip: An empirical study of clip for end to end video clip retrieval and captioning. Neurocomputing **508**, 293–304 (2022)
46. Maji, S., Rahtu, E., Kannala, J., Blaschko, M.B., Vedaldi, A.: Fine-grained visual classification of aircraft (2013)
47. Mu, N., Kirillov, A., Wagner, D.A., Xie, S.: SLIP: self-supervision meets language-image pre-training. In: Eur. Conf. Comput. Vis. vol. 13686, pp. 529–544 (2022)
48. Nilsback, M., Zisserman, A.: Automated flower classification over a large number of classes. In: ICVGIP. pp. 722–729 (2008)
49. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Howes, R., Huang, P.Y., Xu, H., Sharma, V., Li, S.W., Galuba, W., Rabbat, M., Assran, M., Ballas, N., Synnaeve, G., Misra, I., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: DINOv2: Learning robust visual features without supervision (2023)
50. Ordonez, V., Kulkarni, G., Berg, T.: Im2text: Describing images using 1 million captioned photographs. Adv. Neural Inform. Process. Syst. **24** (2011)
51. Parkhi, O.M., Vedaldi, A., Zisserman, A., Jawahar, C.: Cats and dogs. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 3498–3505 (2012)

52. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: PyTorch: An imperative style, high-performance deep learning library. In: Adv. Neural Inform. Process. Syst. pp. 8024–8035 (2019)
53. Radenovic, F., Dubey, A., Kadian, A., Mihaylov, T., Vandenhende, S., Patel, Y., Wen, Y., Ramanathan, V., Mahajan, D.: Filtering, distillation, and hard negatives for vision-language pre-training (2023)
54. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: Int. Conf. Mach. Learn. pp. 8748–8763. PMLR (2021)
55. Recht, B., Roelofs, R., Schmidt, L., Shankar, V.: Do ImageNet classifiers generalize to ImageNet? In: Int. Conf. Mach. Learn. vol. 97, pp. 5389–5400 (2019)
56. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C.W., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., Schramowski, P., Kundurthy, S.R., Crowson, K., Schmidt, L., Kaczmarczyk, R., Jitsev, J.: LAION-5b: An open large-scale dataset for training next generation image-text models. In: Adv. Neural Inform. Process. Syst. (2022)
57. Schuhmann, C., Vencu, R., Beaumont, R., Kaczmarczyk, R., Mullis, C., Katta, A., Coombes, T., Jitsev, J., Komatsuzaki, A.: LAION-400M: Open dataset of CLIP-filtered 400 million image-text pairs (2021)
58. Sharma, P., Ding, N., Goodman, S., Soricut, R.: Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. pp. 2556–2565 (2018)
59. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: Int. Conf. Learn. Represent. (2014)
60. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results **30** (2017)
61. Thomee, B., Shamma, D.A., Friedland, G., Elizalde, B., Ni, K., Poland, D., Borth, D., Li, L.J.: YFCC100M: the new data in multimedia research. Commun. ACM **59**(2), 64–73 (2016)
62. Wang, Z., Wu, Z., Agarwal, D., Sun, J.: MedCLIP: Contrastive learning from unpaired medical images and text (2022)
63. Wei, J.W., Zou, K.: EDA: easy data augmentation techniques for boosting performance on text classification tasks. In: EMNLP-IJCNLP. pp. 6381–6387 (2019)
64. Xiao, J., Ehinger, K.A., Hays, J., Torralba, A., Oliva, A.: SUN database: Exploring a large collection of scene categories. Int. J. Comput. Vis. **119**(1), 3–22 (2016)
65. Xie, C.W., Sun, S., Xiong, X., Zheng, Y., Zhao, D., Zhou, J.: RA-CLIP: Retrieval augmented contrastive language-image pre-training. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 19265–19274 (2023)
66. Xu, H., Xie, S., Huang, P.Y., Yu, L., Howes, R., Ghosh, G., Zettlemoyer, L., Feichtenhofer, C.: CiT: Curation in training for effective vision-language data. In: Int. Conf. Comput. Vis. pp. 15180–15189 (2023)
67. Yao, L., Huang, R., Hou, L., Lu, G., Niu, M., Xu, H., Liang, X., Li, Z., Jiang, X., Xu, C.: FILIP: Fine-grained interactive language-image pre-training. In: Int. Conf. Learn. Represent. (2022)
68. You, H., Zhou, L., Xiao, B., Codella, N., Cheng, Y., Xu, R., Chang, S.F., Yuan, L.: Learning visual representation from modality-shared contrastive language-image pre-training. In: Eur. Conf. Comput. Vis. pp. 69–87. Springer (2022)

69. You, Y., Li, J., Reddi, S., Hseu, J., Kumar, S., Bhojanapalli, S., Song, X., Demmel, J., Keutzer, K., Hsieh, C.J.: Large batch optimization for deep learning: Training BERT in 76 minutes. In: Int. Conf. Learn. Represent. (2020)
70. Young, P., Lai, A., Hodosh, M., Hockenmaier, J.: From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. Trans. Assoc. Comput. Linguistics **2**, 67–78 (2014)
71. Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., Wu, Y.: CoCa: Contrastive captioners are image-text foundation models. Transactions on Machine Learning Research (2022)
72. Yuan, L., Chen, D., Chen, Y., Codella, N., Dai, X., Gao, J., Hu, H., Huang, X., Li, B., Li, C., Liu, C., Liu, M., Liu, Z., Lu, Y., Shi, Y., Wang, L., Wang, J., Xiao, B., Xiao, Z., Yang, J., Zeng, M., Zhou, L., Zhang, P.: Florence: A new foundation model for computer vision (2021)
73. Zhao, L., Zheng, K., Zheng, Y., Zhao, D., Zhou, J.: RLEG: Vision-language representation learning with diffusion-based embedding generation. In: Int. Conf. Mach. Learn. pp. 42247–42258 (2023)
74. Zhou, J., Wei, C., Wang, H., Shen, W., Xie, C., Yuille, A., Kong, T.: iBOT: Image BERT pre-training with online tokenizer. Int. Conf. Learn. Represent. (2022)