This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Misalign, Contrast then Distill: Rethinking Misalignments in Language-Image Pretraining

Bumsoo Kim* Yeonsik Jo Jinhyung Kim Seunghwan Kim
LG AI Research

Abstract

Contrastive Language-Image Pretraining has emerged as a prominent approach for training vision and text encoders with uncurated image-text pairs from the web. To enhance data-efficiency, recent efforts have introduced additional supervision terms that involve random-augmented views of the image. However, since the image augmentation process is unaware of its text counterpart, this procedure could cause various degrees of image-text misalignments during training. Prior methods either disregarded this discrepancy or introduced external models to mitigate the impact of misalignments during training. In contrast, we propose a novel metric learning approach that capitalizes on these misalignments as an additional training source, which we term "Misalign, Contrast then Distill (MCD)". Unlike previous methods that treat augmented images and their text counterparts as simple positive pairs, MCD predicts the continuous scales of misalignment caused by the augmentation. Our extensive experimental results show that our proposed MCD achieves state-of-the-art transferability in multiple classification and retrieval downstream datasets.

1. Introduction

Recent advances in deep learning have shown that image representations trained with large-scale uncurated natural language supervision shows powerful transferability to various downstream tasks [14, 32]. A predominant paradigm in vision–language pre-training is to use a simple contrastive loss that makes the embedding of an image and its matching text description (positive pair) more similar to each other than other arbitrary image–text pairs (negative pairs) [29]. To achieve a more data-efficient training, following works actively capitalized on image random augmentation by: (i) joining language–image pretraining objectives with vision self-supervision terms [3, 5] between the augmented images [27, 23, 18] and (ii) involving more pairs of positive/negative supervisions between the augmented images and their original text description [23, 18].

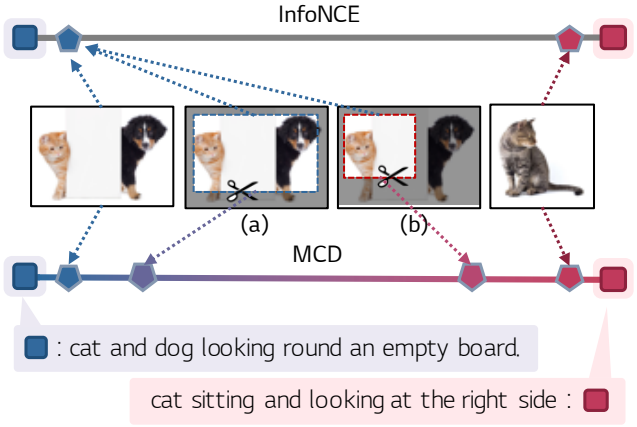

Figure 1. Conceptual illustration of contrastive language–image objectives of previous works (*i.e.*, InfoNCE) and our MCD for (a) augmentation that doesn't harm the correspondence with its description and (b) augmentation that does. Previous works either disregard these misalignments [23] or leverage external models [18, 9] to mitigate their impact. On the other hand, MCD uses the continuous degree of misalignments caused by random image augmentation as a useful source for training various levels of alignments between images and their text descriptions.

However, since the random image augmentation process is unaware of its corresponding text, it often results in the augmented image view to be *misaligned* with its description (see (b) in Fig.1). These misalignments behave as noisy training signals for the contrastive loss in language–image pretraining, thus causing performance degradation if not properly attended [27]. To mitigate this issue, recent works have used additional augmentation embeddings [18] or heavy external off-the-shelf object detectors and summary extractors [9] to match the alignment during training. Though being straightforward and showing strong performance, these works are limited in that they add unnecessary burden in both training and inference.

Based on this observation, we start with a simple question: "Instead of treating misalignments as noise to elim-

*Correspondence to: bumsoo.kim@lgresearch.ai

inate, can we rather harness them as a training source for language-image pretraining?". To this end, we propose MCD (*i.e.*, Misalign, Contrast then Distill), a novel training framework that leverages the various levels of misalignments between random augmented images and its text description during training.

MCD consists of three steps (see Fig 2 for an overview illustration of MCD): First, we conduct random augmentation on the image that causes various levels of misalignments (or not at all) with its text counterpart (**Misalign**). Then, we project all the participants (image, text, and augmented image) into an unified multimodal space, and learn the distance between all the image–text pairs with a contrastive objective (**Contrast**). Finally, we use a teacher-student network where the student learns from the "soft" distance between the text–original image (*i.e.*, $D(\bar{I}, T)$ in Fig 2) and text–augmented image (*i.e.*, $D(\bar{I}', T)$ in Fig 2) of the momentum teacher with a log-ratio loss (**Distill**). continuous scale of misalignment to the student model, enabling the student to learn from the various levels of misalignment that occur from the random augmentation during training time.

Our contribution of this paper is threefold:

- We propose MCD, a novel training framework where we learn the continuous level of misalignment as a source for contrastive language-image pretraining.
- Our MCD outperforms state-of-the-art models across various single/multi-modal downstream datasets without adding additional parameters for inference or using external models to force the alignment.
- We propose three distillation strategies leveraging misalignments: i) misalignment between positive pairs, ii) misalignment between negative pairs, and iii) misalignment between noisy pairs. Extensive experiments show that all three strategies positively contributes to our final performance.

2. Related Work

Here, we present brief review of multi-modal representation learning, especially vision-language pre-training.

2.1. Vision-language Pre-training

Vision-Language Pre-training (VLP) trains a multi-modal model to learn joint representation of visual and textual information that can be transfer to various vision–language downstream tasks. The success of VLP primarily relies on large-scale datasets which contains images and their corresponding descriptions, enabling the model to understand the semantic relationship between the pairs. VLP includes two different group of models: 1) single-stream models [20, 6, 13, 22, 21, 25, 36] which process an image and its associated text information in a shared backbone network and 2) dual-stream models [14, 32] which has two independent backbone networks for processing each modality. In this work, we focuses on Contrastive Language Image Pretraining (CLIP [32]), which is a type of dual-stream models trained with image-text contrastive loss where the image and its matching text description in the dataset as a positive pair and other unrelated pairs in a batch as negative pairs.

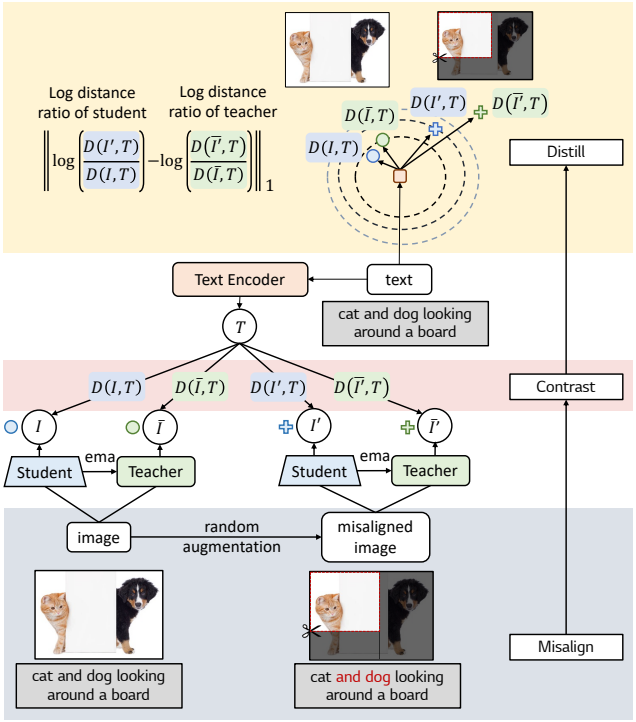

Figure 2. Overview of MCD. Our MCD consist of three steps: (i) text-agnostic random augmentation of the image that causes various levels of misalignment with the corresponding text, (ii) learning the distance between the image/augmented image and text with contrastive objectives then (iii) distill the log-ratio of the image-text distance between the original and the augmented image.

2.2. Misalignments between Image-Text Pairs

There are two different sources that cause misalignments in image-text pairs for VLP: misalignment that naturally occurs in image-text paired datasets, and misalignments caused by random image augmentation.

Misalignment in Image-Text Pairs. Large scale image-text paired datasets for VLP are usually collected from the web thus can contain uncurated and noisy pairs which have weak relations. This inevitably incurs misalignment between positive image-text pairs in the dataset misleading naïve contrastive language image objective. Previous

studies [19, 24, 1] have attempted to address the problem by knowledge distillation [11] of soft image-text alignment matrix from momentum teacher network to the student network via KL divergence loss. Our proposed method stands apart from previous approaches due to its element-wise log-ratio loss for distillation. Element-wise loss lessens the dependence on training hyperparameters like batch size and temperature for KL loss. Furthermore, it enables the model to harness the various levels of individual misalignments of each sample from random augmentation or label noise.

Misalignment by Augmentations. View-based self-supervised learning, in which models are trained to represent views or augmentations of the same image similarly, has yielded strong results across a variety of different formulations. Consecutive work of CLIP (i.e., SLIP [27]) initially introduced supervision between random augmented image views (*e.g.*, cropping, gray-scale, jittering, gaussian blur, horizontal flipping, etc.). Since only image-image supervision was given in a separately embedded space, augmentation was not a factor for misalignment. To include more positive/negative pairs for the image-text contrastive objective, following works [23, 18] also introduced InfoNCE loss between the augmented views and the text pairs. However, since the text description is unaware of the random augmentations, there is a high chance that misalignments occur during training. SLIP shows that naïve application of contrastive loss to these misaligned pairs results in suboptimal performance. Previous work have either ignored this misalignment [23] or addressed this issue with an additional encoder that encodes the one-hot information of which augmentation has been applied during training [18]. PyramidCLIP [9] addressed this issue with external off-the-shelf object detectors and summary extractors. HiCLIP [10] captured the hierarchical nature of high-level on unsupervised manner with tree Transformer [39]. MCD incorporate the misalignment information as a source of training via novel log-ratio loss without introducing any external module.

3. Preliminary

In our preliminary, we revisit the basic form of Contrastive Language-Image Pretraining (*i.e.*, CLIP [32]). CLIP features a dual-encoder architecture where the image encoder f_I and text encoder f_T are jointly trained with a contrastive objective $\mathcal{L}_{\text{CLIP}}$.

InfoNCE Loss Given N image-text pairs $\{(x_i^I, x_i^T)\}_{i=1}^N$, we define a similarity matrix S whose i-th row and the j-th column is the cosine similarity between the projected representations of the i-th text T_i and the j-th image I_j (*i.e.*, $T_i = f_T(x_i^T)$, $I_j = f_I(x_j^I)$), written as:

$$S_{ij} = \text{sim}(T_i, I_j), \tag{1}$$

where $\text{sim}(\cdot,\cdot)$ is cosine similarity. In CLIP [32], the encoded image features I and text features T are projected to the same dimension where the embeddings for matching image-text pairs are pulled together while embeddings for arbitrary pairs are pushed apart with the InfoNCE loss [29]. Given the similarity matrix S, the InfoNCE loss $\mathcal{L}_N$ is rewritten as:

$$\mathcal{L}_N(S) = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp\left(S_{ii}/\tau\right)}{\sum_{j=1}^{N}\exp\left(S_{ij}/\tau\right)}, \tag{2}$$

where τ is a learnable temperature variable. The overall loss of clip $\mathcal{L}_{\text{CLIP}}$ is written as:

$$\mathcal{L}_{\text{CLIP}} = \frac{1}{2}\left(\mathcal{L}_N(S) + \mathcal{L}_N(S^T)\right). \tag{3}$$

Augmentations and Misalignment. To include more positive/negative pairs for the contrastive objective, following works [23, 18] introduced InfoNCE loss between the random augmented image views and the text description for the original image. Let $\mathcal{A}$ a function for random image augmentation that includes randomly applying cropping, gray-scale, jittering, gaussian blur, horizontal flipping following SimCLR [3]. $I_{j'}$ is the encoded image feature of augmented view of j-th image(*e.g.* $I_{j'} = f_I(\mathcal{A}(x_j^I))$). Let $S'_{ij} = \text{sim}(T_i, I_{j'})$ represent the similarity between the augmented view of j-th image and i-th text. We denote the matrix of these similarity as S' for concise notation. Then, InfoNCE loss between the augmented image and the text is

$$\mathcal{L}'_{\text{CLIP}} = \frac{1}{2}\big(\mathcal{L}_N(S') + \mathcal{L}_N(S'^T)\big). \tag{4}$$

As the random augmentation function $\mathcal{A}$ is independent to the corresponding text description, the augmented view $I_{j'}$ is likely to exhibit misalignment with text T_i. This hypothesis is consistent with the finding of SLIP [27], which demonstrated that introducing augmentation (particularly resize crop, and flip) to CLIP actually resulted in a performance decrease. Previous works [27, 23] have sidestepped the utilization of augmented view in CLIP by substituting infoNCE with self-supervised learning loss(*e.g.* SimCLR [3]) between images. These approaches have limitation in fully capturing the essence of multi-modal learning.

4. Method

In this section, we introduce MCD (Misalign, Contrast then Distill), a novel training framework for language–image pretraining using misalignments as *continuous* labels for learning the distance between image–text pairs.

4.1. Misalign

The first step of MCD is to apply text-agnostic random image augmentations (*e.g.*, random crop, random flip, grayscale, etc.) to create various levels of misalignments between the images and its description. Here, we elaborate on three distinct scenarios of misalignment that can arise during the augmentation process, hindering the contrastive loss to learn proper distance between image–text pairs: (i) Text-agnostic random augmentation can cause misalignments in positive image-text pairs. (ii) The random-augmentation can mistakenly cause positive signals between negative pairs. (iii) Misalignments can already exist innately within the original image-text pair. A detailed illustration of each cases are provided in Fig 3.

4.2. Contrast

In the second step, MCD initially learns the distance metric between image–text pairs with a contrastive objective. Motivated by [18], we project both image and text modalities to a unified space and use all the positive pairs and negative pairs of both modalities. For an i-th embedding z_i in a batch of embeddings $\{z_i\}_{i=1}^{3N}$ that includes N image samples, N text samples, and N random augmented image samples, let $\mathcal{P}_i$ and $\mathcal{N}_i$ each denote the set of all positive sample indices of the i-th sample including i itself and the set of all negative sample indices of the i-th sample. Then, the contrastive loss for the multiple positives and multiple negatives for sample i can be written as

$$\mathcal{L}_i^C = \mathbb{E}_{p\in\mathcal{P}_i}\left[-\log\frac{\text{sim}(z_i,z_p)}{\text{sim}(z_i,z_p)+\sum_{n\in\mathcal{N}_i}\text{sim}(z_i,z_n)}\right]. \quad (5)$$

However, without encoding augmentation information the image-text contrastive loss is prone to the three aforementioned issues. We address these three issues with a teacher-student model where the continuous distance between the image–text and augmented image–text of the teacher model is distilled to the student model.

4.3. Distill

Knowledge distillation, introduced by Hinton et al. [11], is a learning paradigm where we train the student network to mimic the "soft" labels predicted from the teacher network. In MCD, we train the student network with the continuous image–text distance predicted by the teacher network. Psuedocode for the distillation in MCD is provided in Algorithm 1.

Log-Ratio Loss for Image–Text Distance. Given a student f_I and a momentum teacher $\bar{f}_I$, we propose to use log-ratio loss [15] on image-text similarities that aims to approximate the ratio of similarity distances by the ratio of

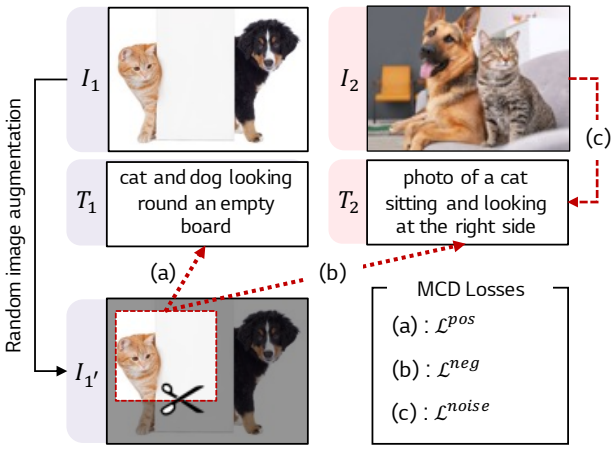

Figure 3. MCD loss for various case of misalignments that occur when applying random image augmentation in language–image pretraining with. In this paper, we elaborate three scenarios: (a) misalignment caused by augmenting the image of an original positive pair, which is addressed by $\mathcal{L}^{\text{pos}}$ (b) augmentation that mistakenly cause positive alignment between negative pairs (c) misalignment that already exists within the dataset.

image-text misalignments in the learned embedding space. Specifically, we define the loss function as

$$\ell_{lr}(i,j;\alpha,\beta) = \left\|\log\frac{D(I_\alpha,T_\beta)}{D(I_i,T_j)} - \log\frac{D(\bar{I}_\alpha,T_\beta)}{D(\bar{I}_i,T_j)}\right\|_1, \quad (6)$$

where $\|\cdot\|_1$ is ℓ_1 loss, $D(\cdot,\cdot)$ is distance function. We utilize Euclidean distance between the projected representations as distance function. Since the embedded vectors comprising the image $I_i, \bar{I}_i$ and text T_j are L2-normalized, Euclidean distance operates as a proxy for cosine similarity $D(I_i,T_j) = 2(1-S_{ij})$. This log-ratio loss approximates the degree of misalignment, which is measured as the ratio of two image-text pairs. By leveraging this measure, we aim to promote a coherent and continuous embedding space. Accordingly, our student encoder is trained under the guidance of the momentum teacher with the incorporation of this degree of misalignment. By establishing these pairs of log-ratio, we enable the handling of diverse forms of misalignment. We present three distinct index setups that correspond to different types of misalignment.

Misalignment in Positive pairs. First, we define distance pair for misalignment between the original image-text pair. Let i' denote the index of augmented image sample. On Eq.(5), i'-th image sample and i-th text sample serve as positive pair. However, random augmentation can occasionally transform positive pair into negative pair. To account for such transformations, we intend to utilize the log-ratio between original pairs and augmented pairs. This allows us to

Algorithm 1 MCD Pseudocode

```
# fm, fs: image encoders (teacher, student)
# ft: text encoder
# fa: random-augmentation function
# N : batch size

def D(sim): # cosine sim -> L2 distance
    return 2 - 2 * sim + 1e-6

def MCD(img, txt):
    # image, text encodings
    aug = fa(img)

    # normalized projection embeddings
    zm, zs, zt = fm(img), fs(img), ft(txt)
    zam, zas = fm(aug), fs(aug) # misalign

    # distance (contrast)
    di_t, di_s = D(zm @ zt.T), D(zs @ zt.T)
    da_t, da_s = D(zam @ zt.T), D(zas @ zt.T)

    # distill
    pos_l, neg_l, noise_l = 0, 0, 0
    for i in range(N):
        lr_s1 = log(da_s[i,i]/di_s[i,i])
        lr_t1 = log(da_t[i,i]/di_t[i,i])

        # positive pairs
        pos_l += abs(lr_s1 - lr_t1) / N

        for j in range(N):
            if i==j: continue
            lr_s2 = log(da_s[i,j]/di_s[i,i])
            lr_t2 = log(da_t[i,j]/di_t[i,i])

            # negative pairs
            neg_l += abs(lr_s2 - lr_t2)/(N*(N-1))

            lr_s3 = log(di_s[j,j]/di_s[i,i])
            lr_t3 = log(di_t[j,j]/di_t[i,i])

            # noisy pairs
            noise_l += abs(lr_s3 - lr_t3)/(N*(N-1))

    return pos_l + neg_l + noise_l
```

capture these shifts and incorporate them into the learning process effectively.

$$\mathcal{L}^{\text{pos}} = \mathbb{E}_{i=1,\dots,N}\left[\ell_{lr}(i,i;i',i)\right], \tag{7}$$

where i' is the index of augmented i-th sample.

Misalignment in Negative pairs. Augmented images can possess relevance with different text, which would normally be considered negative pairs in Eq.(5). However, the log-ratio obtained by our momentum teacher alleviates our model from mistakenly pushing the embedding of relevant texts away.

$$\mathcal{L}^{\text{neg}} = \mathbb{E}_{\substack{i,j=1,2,\dots,N \\ i\neq j}}\left[\ell_{lr}(i,i;j',i)\right], \tag{8}$$

where j' is the index of augmented j-th sample.

Misalignment in Noisy pairs. Original image-text pairs obtained from the web may contain either noisy images or descriptions. While these pairs are normally treated as positive pairs under contrastive loss, we propose a loss for noisy pairs where the noisy labels are trained to have larger distance than matching image–text pairs.

$$\mathcal{L}^{\text{noisy}} = \mathbb{E}_{\substack{i,j=1,2,\dots,N \\ i\neq j}}\left[\ell_{lr}(i,i;j,j)\right]. \tag{9}$$

Distillation Loss. The full training objective of MCD distillation $\mathcal{L}^D$ is the sum of the three distillation terms, which is written as

$$\mathcal{L}^{\text{D}} = \mathcal{L}^{\text{pos}} + \mathcal{L}^{\text{neg}} + \mathcal{L}^{\text{noisy}}. \tag{10}$$

4.4. Training MCD

In this section, we explain the details of MCD training. The training objective for the text encoder and student image encoder for MCD consists of the three objectives: contrastive loss $\mathcal{L}^C$ in Eq (5) for initial image–text distance learning, distillation loss $\mathcal{L}^D$ for the three misalignment scenarios, and loss for masked language modeling $\mathcal{L}^{\text{MLM}}$. The parameters for the teacher image encoder are momentum updated.

MLM Loss. Following previous work in literature [23, 18, 19], we randomly mask out the input tokens with a probability of 15% and replace them with the special token `[MASK]`[1]. Let p^{msk} and y^{msk} each denote the set of model's predicted probability for the masked tokens and the set of ground-truth vocabulary index for the tokens, respectively. Then, MLM loss is written as:

$$\mathcal{L}^{\text{MLM}} = \mathbb{E}_{p\in p^{\text{msk}}, y\in y^{\text{msk}}}\left[\text{CE}(p,y)\right], \tag{11}$$

where CE denotes Cross Entropy loss.

Momentum Teacher Update. Let $\theta_{f_I}, \theta_{\bar{f}_I}$ be the parameter of the student encoder and momentum teacher, respectively. For the t-th step, we update $\theta^{(t)}_{\bar{f}_I}$ of the momentum teacher according to the following:

$$\theta^{(t)}_{\bar{f}_I} = m\theta^{(t-1)}_{\bar{f}_I} + (1-m)\theta^{(t)}_{f_I}, \tag{12}$$

where m denotes the momentum parameter. We use $m = 0.994$ in our experiments, where m grows in a cosine schedule to 1 at the end of training.

Progressive Distillation. As the training proceeds, InfoNCE loss conflicts with our misalignment loss. InfoNCE

[1]Following BERT, the replacement is done with either the `[MASK]` token (80%), another random token within the dictionary (10%), or left unchanged (10%).

between pair I'_j and T_i forces the embedding to pull regardless of its degree of misalignment. In early stages of training, the model needs to learn how to discriminate positive or negative pairs with a hard label. However, as the training progresses, the log-ratio loss delicately models the distance between the various misalignments occurred by augmentations or innately existing in the original image-text pair. Therefore, we progressively diminish the contribution of InfoNCE loss involving augmented views.

MCD Loss. The final loss for MCD is written as:

$$\mathcal{L}^{\text{MCD}} = \mathcal{L}^{\text{C}} + \alpha \cdot \mathcal{L}^{\text{D}} + \beta \cdot \mathcal{L}^{\text{MLM}}, \tag{13}$$

where $\alpha = 0$ progressively increases on a cosine schedule to 1, and $\beta = 0.2$.

4.5. MCD Inference

MCD is based on a teacher-student network, thus the student f_I and momentum teacher $\bar{f}_I$ is obtained after training. Unlike previous work in literature that leverage the teacher network for inference [37, 19], we use the student network that is trained with both the contrastive loss and the log-ratio loss for image–text distance learning.

5. Experiment

In this section, we provide implementation details and experimental results with our MCD pretrained on two widely used image-text benchmark datasets (*e.g.*, CC3M, YFCC15M) to validate the effectiveness of our proposed MCD on multiple downstream datasets including classification and image–text retrieval.

5.1. Implementation Details and Datasets

For implementation details, our work is built on top of the open-source SLIP codebase [27][2]. For DECLIP [23], we follow the implementation details of the official code release[3]. The performance on GPU-machine runs for CLIP and SLIP follows the exact implementation details upon this codebase. Since MCD features both the momentum teacher image encoder $\bar{f}_I$ and student f_I, we conduct the following experiment section with $\bar{f}_I$ based on empirical results. All of our models are pretrained in 16× A100 GPUs. For CC3M, All models are trained with a ViT-B/16 backbone with a learning rate of 5e-4 and weight decay of 0.5. For YFCC15M, we train the model with ViT-B/32 backbone, batch size 4096, learning rate 1e-3, and weight decay 0.2.

Pretraining Datasets. To validate the effectiveness of MCD, we pretrain MCD on large-scale open-source datasets: YFCC (Yahoo Flickr Creative Commons) 15M [38] and CC (Conceptual Captions) 3M [34].

Downstream Datasets. Following CLIP [32], we evaluate the transferability of pretrained MCD on 11 widely used downstream datasets for classification (*i.e.*, Oxford Pets [30], CIFAR-10, CIFAR-100 [17], SUN397 [41], Food-101 [2], Flowers [28], Cars [16], Caltech-101 [8], Aircraft [26], DTD [7], ImageNet-1k [33]). We also transfer to image–text retrieval tasks on Flickr30K [31] and MS-COCO Captions [4] datasets. The evaluation settings for each dataset are consistent with CLIP as in the open-source implementation[2].

5.2. MCD Pretraining on YFCC15M Dataset

First, we pretrain MCD on YFCC15M and evaluate its transferability in single-modal (*e.g.*, classification) and multi-modal (*e.g.*, image–text retrieval) downstream tasks. We compare the result against other state-of-the-art Contrastive Language-Image Pretraining approaches [32, 27, 23, 18] that utilizes various levels of supervision including vision self-supervision [3, 5], text self-supervision [23], memory queue [23], and augmentation encoding [18]. All models are pretrained with a learning rate 1e-3 for 32 epochs unless mentioned otherwise.

Zero-shot Classification. We evaluate the zero-shot classification performance on 11 downstream datasets for single-modal experiments. Tab. 1 shows both the zero-shot classification and linear probing accuracy of CLIP variants [32, 27, 23, 18] pretrained on YFCC15M dataset and transferred to downstream classification datasets. In test time, the learned text encoder f_T synthesizes a zero-shot linear classifier by embedding the arbitrary categories of the test dataset. As classes are in the form of a single word, we use prompts including the label (e.g., "`a photo of a {label}`") as following CLIP [32]. Our MCD outperforms across a majority of the 11 datasets with a noticeable margin. Note that even without additional augmented-aware network leveraged in UniCLIP [18] or additional supervision terms such as text augmentation [40], masked language modeling [23] and memory queue, our MCD achieves state-of-the-art performance.

Linear Probing. To implement linear probe evaluation, we follow CLIP [32] to train a logistic regression classifier on the frozen visual features extracted by the image encoder f_I. Specifically, we train the logistic regression classifier using L-BFGS algorithm provided by scikit-learn with maximum 1,000 iterations, and report the corresponding metric for each dataset[4]. Parameters for L2 regularization are

[2] https://github.com/facebookresearch/SLIP
[3] https://github.com/Sense-GVT/DeCLIP
[4] https://github.com/facebookresearch/SLIP/blob/main/main_linear.py

Method	Vision Encoder	Oxford Pets	CIFAR-10	CIFAR-100	SUN397	Food-101	Flowers	Cars	Caltech-101	Aircraft	DTD	ImageNet	**Average**
Zero-shot Classification:													
CLIP [32]		19.4	62.3	33.6	40.2	33.7	6.3	2.1	55.4	1.4	16.9	31.3	27.5
SLIP [27]		28.3	72.2	45.3	45.1	44.7	6.8	2.9	65.9	1.9	21.8	38.3	33.9
DeCLIP [23]	ViT-B/32	30.2	72.1	39.7	51.6	46.9	7.1	3.9	70.1	2.5	24.2	41.2	35.4
UniCLIP [18]		32.5	78.6	47.2	50.4	48.7	**8.1**	3.4	73.0	2.8	23.3	42.8	37.3
MCD (Ours)		**40.0**	**80.3**	**49.6**	**55.3**	**54.0**	7.9	**4.5**	**73.2**	**3.0**	**30.5**	**44.7**	**40.2**
Linear Probing:													
CLIP [32]		71.2	89.2	72.1	70.1	71.4	93.2	34.9	84.3	29.7	60.9	61.1	67.1
SLIP [27]		75.4	90.5	75.3	73.5	77.1	96.1	43.0	87.2	34.1	71.1	68.1	71.9
DeCLIP [23]	ViT-B/32	76.5	88.6	71.6	75.9	79.3	96.7	42.6	88.0	32.6	69.1	69.2	71.8
UniCLIP [18]		83.1	92.5	78.2	77.0	81.3	97.1	**49.8**	88.9	36.2	72.8	70.8	75.2
MCD (Ours)		**85.6**	**92.7**	**79.3**	**77.6**	**81.7**	**97.1**	46.9	**89.5**	**36.6**	**74.1**	**71.3**	**75.7**

Table 1. Zero-shot image classification/linear probing performance on 11 downstream datasets with YFCC15M pretrained models. Note that DeCLIP [23] utilizes an external momentum queue while UniCLIP [18] features the augmentation encoder during training.

determined using hyperparameter sweep on the validation sets. Standard cropping and flipping augmentations [35] are used for linear probing. The bottom section of Tab. 1 reports linear classification performances on the 11 downstream datasets. Our proposed approach, MCD, has consistently outperformed previous baseline methods in zero-shot classification across multiple datasets, with only one exception.

Image–Text Retrieval. For multi-modal evaluations, we test both the zero-shot and fine-tuned image–text (and text–image) retrieval on Flickr30k and COCO Captions benchmarks. Image-text pairs are ranked according to their similarity scores. Tab. 2 shows the performance for image–text retrieval tasks of MCD pretrained on YFCC15M dataset. Our MCD outperforms all state-of-the-art baselines across every measure with a considerable margin. By incorporating a log-ratio loss with metric learning characteristics into the CLIP framework, our proposed approach has achieved significant improvements in image-text retrieval performance.

Vision–Language Compositionality. To conduct a thorough analysis of multimodal representation learning, we assess the performance of our model using the SugarCrepe [12] dataset. This dataset serves as a de-biased benchmark specifically designed for evaluating the compositionality aspect of vision-language models. SugarCrepe introduces a set of challenging negative captions for COCO image-text pairs by replacing, swapping, or adding certain concepts to the ground truth caption, and gauge the model's capability to discern the positive from its distractor. Tab. 3 summarizes the result, showcasing the performance of various models pretrained on the YFCC15M dataset. Our findings demonstrate that the MCD exhibits significantly better compositionality compared to prior methods. This improved performance is attributed to leveraging augmentations with proper handling for misalignment arising from these augmentations. While other methods often show improved performance over CLIP baseline, adopting augmentations without meticulous management of misalignment cannot maximize their utility.

5.3. MCD Pretraining on CC3M Dataset

In this section, we compare MCD against other state-of-the-art Contrastive Language-Image Pretraining approaches [32, 27, 23]. All models are pretrained on the CC3M dataset with a learning rate 5e-4 for 40 epochs[5]. Tab. 4 shows the ImageNet zero-shot results of MCD with other CLIP variants. MCD outperforms all CLIP variants without external training sources such as Nearest Neighbor supervision with large memory queues (NNS) or augmentation information during training (AUG). Furthermore, MCD does not require any additional parameters for the SSL projection layer [27, 23] or additional network for augmentation-aware feature embedding [18].

5.4. Ablation Study

This section presents ablation studies to evaluate the contribution of each component in our proposed approach, MCD, towards the final performance. To this end, we pretrain all models on the YFCC15M dataset and evaluate them using zero-shot learning on the Imagenet-1k validation set.

[5]More detailed training configuration will be provided in supplement.

Method	Image-to-text retrieval						Text-to-image retrieval					
	Flickr30k			COCO Captions			Flickr30k			COCO Captions		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
Zero-shot retrieval:												
CLIP [32]	34.9	63.9	75.9	20.8	43.9	55.7	23.4	47.2	58.9	13.0	31.7	42.7
SLIP [27]	47.8	76.5	85.9	27.7	52.6	63.9	32.3	58.7	68.8	18.2	39.2	51.0
DeCLIP [23]	51.4	80.2	88.9	28.3	53.2	64.5	34.3	60.3	70.7	18.4	39.6	51.4
UniCLIP [18]	52.3	81.6	89.0	32.0	57.7	69.2	34.8	62.0	72.0	20.2	43.2	54.4
MCD (Ours)	**57.6**	**82.6**	**91.1**	**32.3**	**58.7**	**71.2**	**36.4**	**64.8**	**74.1**	**20.7**	**43.5**	**55.3**
Fine-tuned retrieval												
CLIP [32]	58.3	84.8	91.5	36.1	65.0	76.4	43.1	71.1	80.3	24.9	51.7	64.1
SLIP [27]	69.6	90.4	95.7	45.0	74.0	83.0	52.1	79.4	86.9	31.6	59.5	71.3
DeCLIP [23]	75.6	93.0	96.6	48.7	77.3	86.2	57.8	83.3	90.3	34.2	63.1	74.6
UniCLIP [18]	78.1	94.9	97.7	54.5	80.9	89.1	61.0	86.0	91.9	38.0	67.2	78.0
MCD (Ours)	**79.3**	**95.2**	**98.0**	**55.6**	**81.2**	**89.5**	**63.1**	**87.2**	**92.3**	**38.2**	**67.4**	**78.5**

Table 2. Zero-shot & Fine-tuned image–text retrieval on the test splits of Flickr30k and COCO Captions with models pre-trained on YFCC15M. ViT-B/32 is used for all setup.

Method	Aug.	Misalign	Replace	Swap	Add
CLIP [32]		N/A	73.6	59.5	69.4
SLIP [27]	✓	N/A	74.7	58.6	69.1
DeCLIP [23]	✓	Disregard	74.5	58.2	66.8
UniCLIP [18]	✓	f_A	75.5	58.4	70.4
MCD (Ours)	✓	$\mathcal{L}_{\text{D}}$	**76.2**	**61.5**	**71.7**

Table 3. Evaluation on SugarCrepe [12]. All models were pre-trained on YFCC15M with ViT-B/32 backbone. Models except CLIP involve random image augmentations (Aug.) with different schemes for dealing with image–text misalignments (Misalign). While previous methods either disregard the issue [23] or introduce an additional augmentation encoder (f_A) [18], MCD manages to harness the misalignment for the distillation loss ($\mathcal{L}_{\text{D}}$).

Method	Encoder	SSL	EXT	Top1(%)
CLIP [32]	ViT-B/16	-	-	19.6
SLIP [27]		SimCLR [3]	-	23.2
DeCLIP [23]		SimSiam [5]	NNS	25.4
UniCLIP [18]		MP-NCE [18]	AUG	27.8
MCD (Ours)		MP-NCE [18]	-	**28.2**

Table 4. ImageNet-1k Top 1 zero shot accuracy with models pre-trained on CC3M dataset. SSL denotes the vision self-supervision term used in each model and EXT denotes external sources involved during training. NNS is nearest neighbor supervision using a separate memory queue, and AUG is the vectorized information of each random augmentation conducted during training.

Specifically, we implement the MP-NCE loss without augmentation encoding, which results in an accuracy of 39.6. Our results in Tab. 5 demonstrate that each loss component in MCD has a positive impact on the final performance, leading to an overall improvement in accuracy. These findings highlight the importance of each component in our proposed approach and validate its effectiveness in improving the zero-shot classification performance. Note that MCD outperforms (b) (*i.e.*, UniCLIP) that explicitly includes the augmentation information during training with only $\mathcal{L}^{\text{pos}}$, showing the effectiveness of harnessing the misalignments that occur during random image augmentation for training.

	$\mathcal{L}^{\text{base}}$	f_A	$\mathcal{L}^{\text{pos}}$	$\mathcal{L}^{\text{neg}}$	$\mathcal{L}^{\text{noisy}}$	Top1 Acc (%)
(a)	✓					39.6
(b)	✓	✓				42.8
(c)	✓		✓			**43.9**
(d)	✓		✓	✓		**44.3**
(e)	✓		✓	✓	✓	**44.7**

Table 5. Ablation study on ImageNet-1k Top 1 zero shot accuracy for vision-language pretraining for each loss components of MCD. All models were pretrained with a ViT-B/32 backbone with a basic contrastive loss $\mathcal{L}^{C}$ in Eq (5) and MLM loss in Eq (11), which we abbreviate as $\mathcal{L}^{\text{base}}$. All $\mathcal{L}^{\text{pos}}$, $\mathcal{L}^{\text{neg}}$, $\mathcal{L}^{\text{noisy}}$ shows a consistent gain in zero-shot performance. f_A denotes the augmentation encoder, making (b) analogous to UniCLIP [18].

6. Conclusion

We propose MCD, a new training strategy for dealing with misalignments occured by random image augmentations under visual–language pretraining. Our novel distillation formulation enables data-efficient training under Contrastive Language-Image Pretraining. Future works will include extending MCD frameworks to other modalities.

References

[1] Alex Andonian, Shixing Chen, and Raffay Hamid. Robust cross-modal representation learning with progressive self-distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[2] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *European conference on computer vision*, pages 446–461. Springer, 2014.

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

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

[5] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15750–15758, 2021.

[6] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer, 2020.

[7] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3606–3613, 2014.

[8] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 conference on computer vision and pattern recognition workshop*, pages 178–178. IEEE, 2004.

[9] Yuting Gao, Jinfeng Liu, Zihan Xu, Jun Zhang, Ke Li, and Chunhua Shen. Pyramidclip: Hierarchical feature alignment for vision-language model pretraining. *arXiv preprint arXiv:2204.14095*, 2022.

[10] Shijie Geng, Jianbo Yuan, Yu Tian, Yuxiao Chen, and Yongfeng Zhang. HiCLIP: Contrastive language-image pretraining with hierarchy-aware attention. In *The Eleventh International Conference on Learning Representations*, 2023.

[11] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[12] Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality. *arXiv preprint arXiv:2306.14610*, 2023.

[13] Haoyang Huang, Yaobo Liang, Nan Duan, Ming Gong, Linjun Shou, Daxin Jiang, and Ming Zhou. Unicoder: A universal language encoder by pre-training with multiple cross-lingual tasks. *arXiv preprint arXiv:1909.00964*, 2019.

[14] 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.

[15] Sungyeon Kim, Minkyo Seo, Ivan Laptev, Minsu Cho, and Suha Kwak. Deep metric learning beyond binary supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2288–2297, 2019.

[16] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 554–561, 2013.

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

[18] Janghyeon Lee, Jongsuk Kim, Hyounguk Shon, Bumsoo Kim, Seung Hwan Kim, Honglak Lee, and Junmo Kim. Uniclip: Unified framework for contrastive language-image pre-training. In *Advances in Neural Information Processing Systems*, 2022.

[19] Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Deepak Gotmare, Shafiq Joty, Caiming Xiong, and Steven Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In *Advances in Neural Information Processing Systems*, 2021.

[20] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.

[21] Wei Li, Can Gao, Guocheng Niu, Xinyan Xiao, Hao Liu, Jiachen Liu, Hua Wu, and Haifeng Wang. Unimo: Towards unified-modal understanding and generation via cross-modal contrastive learning. *arXiv preprint arXiv:2012.15409*, 2020.

[22] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer, 2020.

[23] 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.

[24] Haoyu Lu, Nanyi Fei, Yuqi Huo, Yizhao Gao, Zhiwu Lu, and Ji-Rong Wen. Cots: Collaborative two-stream vision-language pre-training model for cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15671–15680, 2022.

[25] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019.

[26] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.

[27] Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. Slip: Self-supervision meets language-image pre-training. In *Computer Vision–ECCV 2022: 17th European*

Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVI, pages 529–544. Springer, 2022.

[28] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pages 722–729. IEEE, 2008.

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

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

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

[32] 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.

[33] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 2014.

[34] 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.

[35] 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.

[36] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019.

[37] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, 2017.

[38] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016.

[39] Yaushian Wang, Hung-Yi Lee, and Yun-Nung Chen. Tree transformer: Integrating tree structures into self-attention. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1061–1070, Hong Kong, China, 2019. Association for Computational Linguistics.

[40] Jason Wei and Kai Zou. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388, 2019.

[41] Jianxiong Xiao, Krista A Ehinger, James Hays, Antonio Torralba, and Aude Oliva. Sun database: Exploring a large collection of scene categories. *International Journal of Computer Vision*, 119(1):3–22, 2016.