

SaCo Loss: Sample-wise Affinity Consistency for Vision-Language Pre-training

Sitong Wu^{1,*} Haoru Tan^{2,*} Zhuotao Tian³ Yukang Chen¹
Xiaojuan Qi² Jiaya Jia^{1,3,†}

¹CUHK ²HKU ³SmartMore

<https://github.com/dvlab-research/SaCo-Loss>

Abstract

Vision-language pre-training (VLP) aims to learn joint representations of vision and language modalities. The contrastive paradigm is currently dominant in this field. However, we observe a notable misalignment phenomenon, that is, the affinity between samples has an obvious disparity across different modalities, namely “**Affinity Inconsistency Problem**”. Our intuition is that, for a well-aligned model, two images that look similar to each other should have the same level of similarity as their corresponding texts that describe them. In this paper, we first investigate the reason of this inconsistency problem. We discover that the lack of consideration for sample-wise affinity consistency across modalities in existing training objectives is the central cause. To address this problem, we propose a novel loss function, named **Sample-wise affinity Consistency (SaCo)** loss, which is designed to enhance such consistency by minimizing the distance between image embedding similarity and text embedding similarity for any two samples. Our SaCo loss can be easily incorporated into existing vision-language models as an additional loss due to its complementarity for most training objectives. In addition, considering that pre-training from scratch is computationally expensive, we also provide a more efficient way to continuously pre-train on a converged model by integrating our loss. Experimentally, the model trained with our SaCo loss significantly outperforms the baseline on a variety of vision and language tasks.

1. Introduction

Vision-language pre-training (VLP) has shown remarkable success since the emergence of contrastive vision-language learning [55], which aims to promote the paired image and text samples closer while simultaneously pushing unpaired

samples away. The following works achieved further improvements through vision masked modeling [9, 33], fine-grained supervision [23, 53], label smoothing [84], augmentations [32], text refinement [11, 21] and noisy data filtering [14, 67], and text generation supervision [26, 27]. However, during the contrastive training process, we observe an Affinity Inconsistency Problem, that is, the correlation between the sample-wise affinity (similarity) in language modality and visual modality is extremely low. For a highly aligned vision-language embedding space, the similarity between every two images shall be similar to the similarity between their corresponding texts, as both the paired image and text describe the same object or scene. The detailed qualitative (Figure 2) and quantitative (Figure 3a) analysis of this problem are discussed in Sec. 3.2.

The reason is that contrastive pre-training mainly focuses on aligning or pushing away the cross-modal embeddings, while ignoring whether the affinity between samples is consistent in different modalities, see Sec. 3.3 for more analysis. For better clarity, we present an illustration in Figure 1a, which depicts an embedding space with the affinity inconsistency problem. Taking three image-text pairs as an example, the circle and square with the same color denote the embedding of paired image and text, respectively. The blue triangle (\triangle) connecting three circles represents the relationship between images. Similarly, the red triangle (\triangle) connecting three squares reflects the relationship between texts. The shorter side between two images or texts means the greater affinity between them. In Figure 1a(ii), we show the geometric inconsistency between “ \triangle ” and “ \triangle ” in the contrastively trained space, highlighting the disparity in sample-wise affinity between the vision and language modalities.

To address this problem, we propose a simple but effective loss function, named *Sample-wise affinity Consistency (SaCo) loss*, whose core idea is to minimize the disparity between image embedding similarity and text embedding similarity for any two image-text pairs. Intuitively, it is designed to specifically promote the consistency of affinity

*Equal contribution.

†Corresponding author.

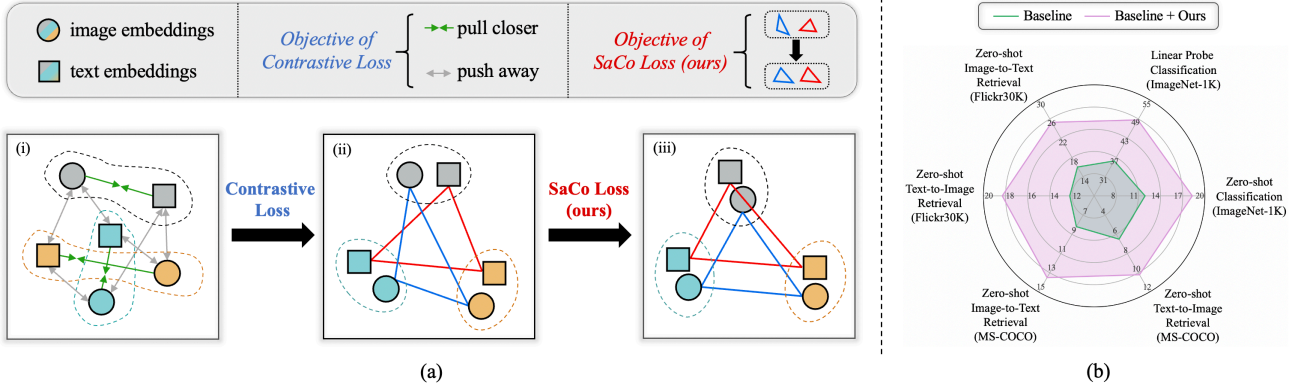


Figure 1. (a) Comparison on the optimization objective of our SaCo loss and contrastive loss. As in (i), contrastive loss aims to pull the embeddings of paired image and text together (\leftrightarrow) while pushing those of unpaired ones apart (\leftrightarrow), optimizing the embedding space to (ii). The blue (\triangle) and red (\triangle) triangles in (ii) represent the image embedding affinity and text embedding affinity, respectively. Our SaCo loss strives for the *geometric consistency* between “ \triangle ” and “ \triangle ”, and optimize the embedding space to (iii) with consistent affinity. (b) Performance comparison on various tasks, where the baseline is CLIP with ViT-B/32 image encoder.

between samples in different modalities. Since our SaCo loss is complementary to most training objectives, it can be seamlessly integrated into existing vision-language models as an additional loss. In practice, pre-training from scratch is computationally expensive. Accordingly, we have also explored another more efficient approach to endow an existing excellent model with the property of cross-modal affinity consistency, which is achieved by performing continuous pre-training for a few epochs with our SaCo loss.

Extensive experiments on various tasks demonstrate the effectiveness of our approach. As shown in Figure 1b, our SaCo loss achieves superior performance compared to the CLIP baseline. It improves Recall@1 performance by 9.3% and 5.3% in zero-shot Image-to-text retrieval on Flickr30K [52] and MS-COCO [38], respectively. Additionally, it enhances Recall@1 performance by 6.1% and 3.7% in zero-shot text-to-image retrieval on the Flickr30K and MS-COCO datasets, respectively. Furthermore, it boosts Top-1 Accuracy performance by approximately 6.4% and 13% in zero-shot and linear prob image classification on the ImageNet-1K dataset. For more evaluations, we have included our experiments on image retrieval and text classification in the supporting materials.

2. Related Work

Vision-language Pre-training. Vision-language pre-training (VLP) is one of the most important advancements in multi-modal learning [2, 20, 55, 68, 80, 82]. It strives to learn general and transferable representations for both visual and linguistic modalities from large-scale data. Earlier works adopted a single-stream structure, using a single transformer to learn both joint image and text representations by concatenating image and text input embed-

dings [4, 24, 28, 30, 31]. Later, the emergence of CLIP [55] opened up the dual-stream structure, which separately encodes image and text with decoupled image and text encoder. CLIP has shown remarkable performance on zero-shot transferability for various downstream tasks by leveraging contrastive learning on large-scale image-text pairs. Subsequently, a series of works continued to improve this contrastive paradigm by introducing new optimization objectives. DeCLIP [32] and SLIP [47] introduced additional self-supervision training objectives in order to improve data efficiency [60]. FLIP [33] and MaskCLIP [9] incorporated the visual masked modeling [13, 88] to enhance local semantics. FILIP [23], LOUPE [25], and VoLTA [53] explored more accurate fine-grained alignment between the two modalities [85, 87, 89]. SoftCLIP [84] and PyramidCLIP [83] relaxed the strict one-to-one constraint to a soft cross-modal alignment. Image-text matching is further integrated into the contrastive paradigm as a complement [1, 26, 27] to predict whether an image-text pair is positive (matched) or negative (unmatched) via binary classification. To facilitate both understanding tasks [5, 10, 13, 17, 36, 37, 75] and generation tasks [90], CoCa [44] and BLIP series [26, 27, 92] combined a text generation objective through language-masked modeling. Different from them, this paper focuses on another objective, targeting the consistent sample-wise affinity across modalities. In addition, our approach is complementary to previous training objectives and therefore can be seamlessly incorporated into existing vision-language models.

Consistency-based Supervision. The consistency-based supervision is highly favored in deep learning. Cycle consistency is commonly used in bidirectional scenarios, such as machine translation [3, 7], unpaired image-to-image translation [18], vision-language generation [29], cross-












Query		The bus leaving english metropolitan borough ...	A vintage double decker bus.	An old bus full of people on a trip	A line of historic buses on display.	Bus is seen in service on the capital's streets.
 <p>Quarter front view of automobile model of a bus seen here.</p>	Image query results					
	Text query results	A general view of the automobile model.	Automobile model on four wheels, with a black lip.	View this image of automobile model.	A front view of an old car.	View this image of automobile model.
						

Figure 2. Qualitative illustration for the “affinity inconsistency problem”. For the query image-text pair, we query its neighbors in a multi-modal dataset based on its image embedding or text embedding, separately. The embeddings are obtained from the official CLIP [55]. However, queries in different modalities return very different results. When querying with image embedding, the top-ranked images (in green shadow) are all related to “bus”. While the top-ranked texts (in blue shadow) are mostly related to “automobile” when querying with image embedding.

modal retrieval [8, 39, 69–71, 73] and domain adaption [81]. Temporal consistency is an essential technique in various video generation and understanding tasks [12, 22, 46, 48, 77, 91]. Similarly, spatial consistency is also an effective prior in computer vision, for example, point cloud registration [78], graph matching [40, 59, 61], distillation [51], representation learning [66, 76] and scene understanding [19, 34, 35, 42, 45, 58, 62–65, 74, 79]. Furthermore, feature consistency, between local and global, can facilitate deep feature clustering [6], and neighbor consistency in the feature or prediction space can improve noisy learning [15, 43]. In this paper, we extend consistency supervision to the field of vision-language representation learning, which introduces a new method for exploiting the inherent consistency among different modalities.

3. Affinity Inconsistency Problem

In this section, we first revisit the current dominant contrastive pre-training paradigm in Sec. 3.1. Then, we experimentally explore the affinity inconsistency problem in Sec. 3.2 and further investigate its reasons in Sec. 3.3.

3.1. Preliminaries

The contrastive language-image loss [55] has been utilized in most popular vision-language models [16, 26, 27, 54, 55, 72, 86] during pre-training. It aims to learn an aligned embedding space for visual and language modalities. Specifically, given a batch of N image-text pairs $\{(\mathbf{x}_i^I, \mathbf{x}_i^T)\}_{i=1}^N$, each image \mathbf{x}_i^I and text \mathbf{x}_i^T are passed through the image encoder and text encoder independently to get their corresponding embedding. After the linear projection and L2 normalization, we obtain the final embed-

dings $\{(\mathbf{I}_i, \mathbf{T}_i)\}_{i=1}^N$ for all the pairs, where the paired image and text $(\mathbf{I}_i, \mathbf{T}_i)$ is treated as the positive pair and the unpaired one $(\mathbf{I}_i, \mathbf{T}_{j(j \neq i)})$ forms the negative pair. The contrastive loss can be formulated as follows:

$$\mathcal{L}_{\text{cont}} = -\frac{1}{N} \sum_{i=1}^N \log \frac{e^{(\mathbf{S}_{ii}/\tau)}}{\sum_{j=1}^N e^{(\mathbf{S}_{ij}/\tau)}} - \frac{1}{N} \sum_{i=1}^N \log \frac{e^{(\mathbf{S}_{ii}/\tau)}}{\sum_{k=1}^N e^{(\mathbf{S}_{ki}/\tau)}}, \quad (1)$$

where $\mathbf{S}_{ij} = \langle \mathbf{I}_i, \mathbf{T}_j \rangle$ is the cross-modal inner-product similarity, and τ is a learnable temperature parameter [55] to control the smoothness of distribution. The contrastive loss minimizes the distance between the embeddings of the positive pair across different modalities and maximizes the distance between the embeddings of the negative pair. Thus, the contrastive loss enables images or texts that share the same semantic information to have similar representations in the feature space.

3.2. Problem Discovery

Contrastive pre-training can basically ensure that images and texts with the same semantic information also have similar representations in the feature space, thereby aligning the embeddings in different modalities. However, we observe a significant misalignment phenomenon. Specifically, given the i -th image-text pair from a multi-modal dataset (such as CC3M [56]), we first utilize a public vision-language model to extract its image embedding \mathbf{I}_i and text embedding \mathbf{T}_i , which are separately used to compute the image embedding similarities $\mathbf{S}_i^I \in \mathbb{R}^N$ and text embedding similarities $\mathbf{S}_i^T \in \mathbb{R}^N$ between the query and all the other samples within the dataset. The calculation is as follows,

$$\mathbf{S}_{i,j}^I = \langle \mathbf{I}_i, \mathbf{I}_j \rangle, \quad \mathbf{S}_{i,j}^T = \langle \mathbf{T}_i, \mathbf{T}_j \rangle, \quad (2)$$

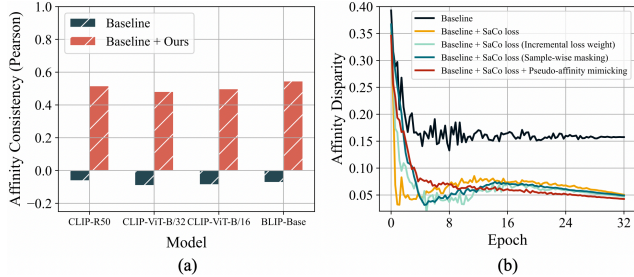


Figure 3. (a) Comparison of the affinity consistency metric defined in Eq.(4). We use the public CLIP [55] and BLIP [26] as baseline models. (b) Comparison of the disparity of affinity across modalities, which is measured by the SaCo loss defined in Eq.(5). We show the fluctuation of this metric during training under different objectives.

where $\langle \cdot, \cdot \rangle$ denotes the inner-product operation.

Qualitative Analysis. We first sort the samples in descending order based on the two similarity vectors. This process yields a sorted sequence of image similarities and a sorted sequence of text similarities. Ideally, these two sequences should be highly consistent, as both the given image and text represent the same object or scene. While, in practice, they exhibit obvious inconsistency. A qualitative example is shown in Figure 2, containing the top-5 entities from both sequences.

Quantitative Analysis. For more quantitative analysis, we utilize the Pearson correlation coefficient defined as follows,

$$\rho(\mathbf{U}, \mathbf{V}) = \frac{\sum_{i=1}^C (\mathbf{U}_i - \bar{\mathbf{U}})(\mathbf{V}_i - \bar{\mathbf{V}})}{\sqrt{\sum_i^C (\mathbf{U}_i - \bar{\mathbf{U}})^2 \sum_i^C (\mathbf{V}_i - \bar{\mathbf{V}})^2}}, \quad (3)$$

where C is the dimensional size of input vectors, $\bar{\mathbf{U}}$ and $\bar{\mathbf{V}}$ indicate the mean value. The Pearson correlation coefficient ranges from -1 to +1, encompassing a spectrum of relationships between variables, ranging from negative correlation to positive correlation. When $\rho = 0$, there is no correlation between the examined variables. Specifically, we iteratively treat each sample in the dataset as a query sample to calculate the image similarities and text similarities and average their Pearson correlation coefficient to obtain the so-called *affinity consistency* metric, that is,

$$R_{\text{affinity-consistency}} = \frac{1}{|D|} \sum_{i=1}^{|D|} \rho(\mathbf{S}_i^I, \mathbf{S}_i^T), \quad (4)$$

where $|D|$ is the dataset size. As shown in Figure 3a, existing popular models (CLIP [55] and BLIP [26]) present quite low correlation in cross-modal similarity compared to their counterparts pre-trained with our approach. Since the similarity querying result essentially reflects the affinity between each candidate sample and the query sample, we refer to such inconsistency as the affinity inconsistency problem.

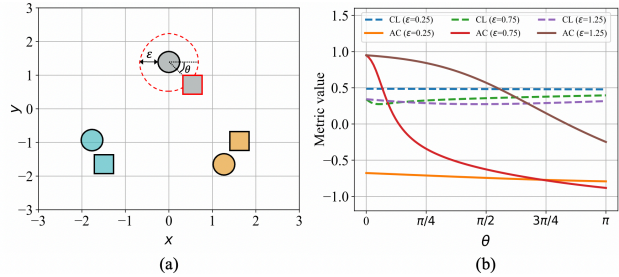


Figure 4. A simulated example to indicate why the dominant contrastive pre-training has the affinity inconsistency problem. (a) illustrates the settings of the simulation. We consider three image-text pairs, each distinguished by a different color and represented by circles and squares. The circles or squares with black edges remain fixed, while only the one square with red edge is allowed to vary by an angle θ along the red dotted circle with a radius of ε from its corresponding image embedding. (b) The results of the simulation demonstrate how the contrastive loss (CL) defined in Eq.(1) and affinity consistency (AC) metric defined in Eq.(4) vary with the angle θ under different radius ε settings.

3.3. Reason Analysis

In this subsection, we investigate why it is difficult for contrastive pre-training to learn consistent cross-modal affinity between samples. First, we monitor the variation of affinity disparity (one minus the averaged affinity consistency) during the standard contrastive pre-training process. As shown in the black curve in Figure 3b, we observe an increasing trend in this metric as training progresses. However, when compared to the training process under our paradigm (red curve), it can be found that the affinity disparity metric of the vanilla contrastive pre-training quickly stagnates at an unsatisfactory level (around 0.15).

Intuitive Simulation. To clearly illustrate this, we also provide an intuitive simulation in Figure 4. As shown in Figure 4a, we take three image-text pairs as an example (represented by different colors). Their image and text embedding are denoted by circles and squares, respectively. Supposing that only one sample’s text embedding (marked with a red edge) is changeable along the circle (denoted as a red dotted circle) of radius ε from its corresponding image embedding, and other embeddings with black edges are fixed. We simulate the training process by varying the radius ε of the circle (also the distance between cross-modal embedding) and the angle θ between the features. In the simulation, we monitor the changes in both the contrastive loss by Eq.(1) and affinity consistency by Eq.(4), as shown in Figure 4b. It can be observed that changes in radius ε significantly affect the contrastive loss, whereas changes in angle θ significantly affect the affinity consistency. This implies that during the training process, the contrastive loss only has a direct impact on the relative distance (similarity) between

cross-modal features, but is less sensitive to their relative angular relationships. However, these angular relationships significantly affect affinity consistency, leading to the affinity inconsistency problem.

4. Methodology

In this section, we first present the design of our proposed SaCo loss in Sec. 4.1, and then introduce two strategies for incorporating our loss function into vision-language pre-training in Sec. 4.2.

4.1. SaCo Loss

To address the affinity inconsistency problem, we propose a simple yet powerful loss function, named sample-wise affinity consistency (SaCo) loss. The idea of our loss is quite intuitive. Specifically, given a batch of N image-text pair samples, they are passed through the corresponding encoder to extract the image embedding $\{\mathbf{I}_i\}_{i=1}^N$ and text embedding $\{\mathbf{T}_i\}_{i=1}^N$, respectively. Note that the embeddings are L2-normalized. First, for the i -th data point, we calculate the image embedding similarity vector \mathbf{S}_i^I and the text embedding similarity vector \mathbf{S}_i^T between the current data and each data within the same batch, check Eq.(2) for details. Then, our SaCo loss can be formulated as follows:

$$\mathcal{L}_{\text{SaCo}} = \sum_{i=1}^N D(\mathbf{S}_i^I, \mathbf{S}_i^T), \quad (5)$$

where $D(\cdot, \cdot)$ denotes the L1 distance between two input vectors as defined in Eq.(3). It measures the consistency degree between the vision-modality affinity and language-modality affinity. Therefore, minimizing this SaCo loss can specifically bridge the gap between vision and language modality on sample-wise affinity. We will show that our SaCo loss is well-compatible with existing vision-language pre-training models via experimentation.

4.2. Vision-language Pre-training with SaCo Loss

Our SaCo loss has broad applicability and can be used as an additional loss function for existing vision-language models. The overall training loss can be expressed as a weighted combination of the original loss functions (such as the contrastive loss in the CLIP model, as shown in Eq. (1)) and our loss. The equation for the total loss is as follows:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{original}} + \alpha \mathcal{L}_{\text{SaCo}}, \quad (6)$$

where α represents the weight coefficient. We then present two strategies for incorporating our SaCo loss into vision-language pre-training, along with detailed designs.

4.2.1 Pre-training from Scratch

Pre-training from scratch is the most commonly used strategy, in which the parameters are randomly initialized before

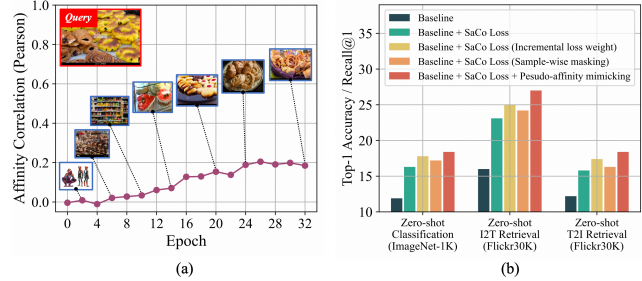


Figure 5. (a) The evolution of affinity accuracy during training. (b) Comparison of different solutions for the instability issue when pre-training from scratch. The pseudo-affinity mimicking strategy performs best. All the experiments are pre-trained on the CC3M dataset and utilize the CLIP with ViT-B/32 image encoder as the baseline model.

training. In practice, we observe some instabilities when pre-training from scratch with our SaCo loss. As shown in the yellow curve in Figure 3b, it first drops sharply, then starts to rise, and finally decreases again in the later training stage. To investigate the reason, we monitor how the accuracy of affinity varies with training. The accuracy is estimated by the correlation between the affinity from different training stages and that from a well-trained model, such as CLIP ViT-L/14. As shown in Figure 5a, in the early training stage, the correlation is quite low and the query results almost have no common visual components with the query image containing “Pastries”. Obviously, the affinity in both modalities is extremely noisy in the early training stage. At this time, it is ineffective to overemphasize the affinity consistency and will hinder the optimization of embedding space. For the aforementioned problem, we design and compare several strategies as follows:

- 1) *Incremental loss weight.* Considering that inaccurate affinity usually appears in the early training stage, we gradually increase the weight α of the SaCo loss in Eq.(6) as the training progresses.
- 2) *Sample-wise masking.* We utilize the similarity between image and text embeddings as a criterion to assess the representation quality. The larger the similarity, the more reliable the representations are. By thresholding the similarity, we obtain a binary mask that indicates whether each sample needs to be subjected to SaCo loss.
- 3) *Pseudo-affinity mimicking.* We allow the training model to mimic a pseudo-affinity predicted by a public well-trained model (e.g., CLIP [55], DINOv2 [49]). Similar to the SaCo loss, this pseudo-affinity mimicking process can be formulated as follows:

$$\mathcal{L}_{\text{mimic}} = \sum_{i=1}^N D(\mathbf{S}_i^I, \tilde{\mathbf{S}}_i^I), \quad (7)$$

where $\tilde{\mathbf{S}}_i^I$ is the predicted pseudo-affinity, and $D(\cdot, \cdot)$ denotes L1 distance. This objective is also integrated via

weighted summation, and the total loss is an extension to Eq.(6), that is,

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{original}} + \alpha\mathcal{L}_{\text{SaCo}} + \beta\mathcal{L}_{\text{mimic}}, \quad (8)$$

where β is the weight for pseudo-affinity mimicking.

Experimentally, we find that the third solution is the most effective choice (see Figure 5b). Note that this mimicking is better to be only performed in the visual modality (detailed in Sec. 5.4 and Table 5). Please refer to supplementary materials for more motivations, details, analysis, and other potential solutions.

4.2.2 Continue Pre-training

Given a well-trained vision-language model (VLM), one can also continuously pre-train the well-trained model by incorporating our SaCo loss. In this case, we empirically find that it is no longer necessary to mimic the pseudo-affinity for training stability, as the affinity of a well-trained model has already been optimized to a good state. In addition, under this strategy, it is not indispensable to use the original pre-training dataset of the well-trained VLM. As shown in Table 3, even a relatively small public dataset can lead to significant improvement. Therefore, it is beneficial for models trained with private data and efficient training.

5. Experiments

We first briefly describe the experimental setups in Sec. 5.1. Then, we analyze the performance on various vision-language tasks in Sec. 5.2 and 5.3. Ablation studies are conducted in Sec. 5.4 to investigate the effect of each key design and hyper-parameter.

5.1. Experimental Setup

Datasets. For pre-training, the experiments are conducted on two open-source image-text datasets at different scales: CC3M [56] and YFCC15M [32]. We perform evaluations across a wide range of tasks. The datasets utilized for each task are detailed in the supplementary materials.

Baseline Models. In our experiments, the main baseline model is the CLIP [55] with three kinds of image encoders: ResNet-50 [55], ViT-B/32 [10], and ViT-B/16 [10].

Pre-train Settings. The experiments are conducted on 16 NVIDIA V100 GPUs, and implemented with PyTorch [50]. Details are listed in supplementary materials.

5.2. Comparison on Pre-training from Scratch

Zero-shot Classification. Zero-shot classification [57] requires a model to classify data it has never been explicitly trained on. Table 1 compares the zero-shot ImageNet-1K classification accuracy of the models pre-trained from scratch on datasets with different scales. When pre-training

Pre-train Dataset	Method	Image Encoder	ImageNet-1K	
			Top-1 Acc.	Top-5 Acc.
CC3M	CLIP	R50	17.9	36.3
	CLIP + Ours	R50	22.5 (+4.6)	41.3 (+5.0)
	CLIP	ViT-B/32	11.9	26.2
	CLIP + Ours	ViT-B/32	18.3 (+6.4)	35.1 (+8.9)
	CLIP	ViT-B/16	16.6	33.1
	CLIP + Ours	ViT-B/16	21.8 (+5.2)	40.0 (+6.9)
YFCC15M	CLIP	R50	37.2	62.1
	CLIP + Ours	R50	42.0 (+4.8)	67.6 (+5.5)
	CLIP	ViT-B/32	31.2	55.7
	CLIP + Ours	ViT-B/32	38.0 (+6.8)	63.2 (+7.5)
	CLIP	ViT-B/16	37.6	63.5
	CLIP + Ours	ViT-B/16	43.5 (+5.9)	69.3 (+5.8)

Table 1. Zero-shot ImageNet-1K classification results for the models pre-trained from scratch. ‘‘Acc.’’ is the short for ‘‘Accuracy’’.

on CC3M, our approach can bring consistent performance improvement to all the CLIP variants, for example, +4.6%, +5.2%, and +6.4% accuracy gain for R50, ViT-B/16, and ViT-B/32 respectively. The model with 15M pre-trained data still shows similar improvement from +4.8% to +6.8% top-1 accuracy across all variants.

Zero-shot Image-Text Retrieval. The zero-shot retrieval experiments include image-to-text (I2T) retrieval and zero-shot text-to-image (T2I) retrieval. The recall performance on these two tasks is summarized in Table 2, considering two widely-used benchmarks. Strikingly, pre-training with our loss yields a significant improvement over the baseline by a large margin. Specifically, for the Flickr30K [52] dataset, we achieve a remarkable improvement ranging from +8.0% to +12.8% in Recall@1 for I2T retrieval, and +6.1% to +7.9% in Recall@1 for T2I retrieval, across different baseline variants. For the more challenging MS-COCO [38] dataset, our loss leads to an improvement of approximately +5% in Recall@1 for I2T retrieval, and a gain of +2.7% to +3.7% in Recall@1 for T2I retrieval. When considering the top-10 retrieval results (measured by Recall@10), the improvement more than doubles, achieving +8.0% to +11.5% for I2T retrieval, and +4.3% to +10.3% for T2I retrieval.

When pre-training on a larger dataset (YFCC15M [32]), our loss continues to achieve sustained improvements on the MS-COCO dataset. We observe an approximate increase of +3% to +5% in Recall@1 and +4% to +7% in Recall@5 for both I2T and T2I retrieval. On the Flickr30K dataset, although the improvement margin slightly narrows compared to that of the CC3M pre-trained model, it is still substantial. We observe an improvement of approximately +5% in Recall@1 for I2T retrieval and +4.8% to +6.4% in Recall@1 for T2I retrieval. It is worth highlighting the significant performance gain achieved with our loss on both datasets.

Pre-train Dataset	Method	Image Encoder	Flickr30K (1K test set)		MS-COCO (5K test set)	
			Image → Text R@1 / R@5 / R@10	Text → Image R@1 / R@5 / R@10	Image → Text R@1 / R@5 / R@10	Text → Image R@1 / R@5 / R@10
CC3M	CLIP	R50	29.2 / 58.5 / 70.5	24.8 / 51.7 / 63.2	16.4 / 38.2 / 50.2	13.3 / 32.4 / 43.5
	CLIP + Ours	R50	42.0 / 72.1 / 80.1 (+12.8 / +13.5 / +9.6)	31.8 / 57.8 / 67.9 (+7.0 / +6.1 / +4.7)	21.1 / 45.3 / 58.1 (+4.7 / +7.1 / +7.9)	16.0 / 36.5 / 47.8 (+2.7 / +4.1 / +4.3)
	CLIP	ViT-B/32	16.0 / 39.1 / 51.7	12.2 / 30.1 / 40.3	8.2 / 22.5 / 32.0	6.5 / 18.2 / 25.8
	CLIP + Ours	ViT-B/32	25.3 / 51.1 / 63.8 (+9.3 / +12.0 / +12.1)	18.3 / 40.0 / 50.5 (+6.1 / +9.9 / +10.2)	13.5 / 32.5 / 43.5 (+5.3 / +10.0 / +11.5)	10.2 / 25.8 / 36.1 (+3.7 / +7.6 / +10.3)
	CLIP	ViT-B/16	26.7 / 54.4 / 65.9	18.8 / 41.0 / 52.4	13.2 / 31.5 / 43.2	10.0 / 25.6 / 35.8
	CLIP + Ours	ViT-B/16	34.7 / 65.8 / 76.6 (+8.0 / +11.4 / +10.7)	26.7 / 51.4 / 62.0 (+7.9 / +10.4 / +9.6)	17.2 / 39.6 / 51.2 (+4.0 / +8.1 / +8.0)	13.3 / 32.1 / 43.1 (+3.3 / +6.5 / +7.3)
YFCC15M	CLIP	R50	54.0 / 81.3 / 87.7	35.6 / 63.5 / 74.0	27.8 / 51.9 / 63.1	16.3 / 37.1 / 48.9
	CLIP + Ours	R50	59.1 / 84.6 / 91.0 (+5.1 / +3.3 / +3.3)	40.4 / 66.8 / 76.1 (+4.8 / +3.3 / +2.1)	32.7 / 58.8 / 70.4 (+4.9 / +6.9 / +7.3)	20.9 / 43.8 / 55.6 (+4.6 / +6.7 / +6.7)
	CLIP	ViT-B/32	42.1 / 68.9 / 78.8	25.3 / 50.1 / 61.9	23.1 / 46.0 / 57.8	13.8 / 32.6 / 43.7
	CLIP + Ours	ViT-B/32	47.2 / 75.6 / 85.6 (+5.1 / +6.7 / +6.8)	31.7 / 57.7 / 67.7 (+6.4 / +7.6 / +5.8)	27.2 / 53.4 / 65.9 (+4.1 / +7.4 / +8.1)	17.9 / 39.3 / 50.6 (+4.1 / +6.7 / +6.9)
	CLIP	ViT-B/16	53.5 / 80.5 / 89.1	34.5 / 61.6 / 72.8	30.3 / 55.6 / 66.7	18.0 / 40.4 / 51.8
	CLIP + Ours	ViT-B/16	58.9 / 85.7 / 91.9 (+5.4 / +5.2 / +2.8)	40.5 / 67.9 / 77.0 (+6.0 / +6.3 / +4.2)	33.4 / 60.1 / 71.6 (+3.1 / +4.5 / +4.9)	22.7 / 45.4 / 56.7 (+4.7 / +5.0 / +4.9)

Table 2. Zero-shot image-text retrieval results. All the models are pre-trained from scratch. “R@k” is short for “Recall@k”.

Method	Image Encoder	Flickr30K (1K test set)				MS-COCO (5K test set)				ImageNet-1K Top-1 Acc.
		Image → Text		Text → Image		Image → Text		Text → Image		
		R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5	
CLIP	ViT-B/32	77.8	94.1	57.0	82.8	49.6	73.6	29.3	54.7	63.0
CLIP + Ours	ViT-B/32	81.7 (+3.9)	95.5 (+1.4)	64.9 (+7.9)	88.1 (+5.3)	54.6 (+5.0)	77.6 (+4.0)	36.1 (+6.8)	62.0 (+7.3)	64.0 (+1.0)
CLIP	ViT-B/16	81.7	96.0	60.8	85.1	51.8	75.7	32.0	57.2	68.2
CLIP + Ours	ViT-B/16	85.5 (+3.8)	96.5 (+0.5)	69.1 (+8.3)	90.1 (+5.0)	57.8 (+6.0)	80.0 (+4.3)	39.8 (+7.8)	64.7 (+7.5)	69.3 (+1.1)

Table 3. Performance of the continuously pre-trained models on zero-shot classification and image-text retrieval. We perform continue pre-training on the baseline CLIP model published by OpenAI. “R@k” is short for “Recall@k”.

Summary and Analysis. Interestingly, we notice about 1~2 times the performance gain on I2T retrieval than T2I retrieval, which indicates that our loss brings a more prominent improvement for the text embedding space than image embedding space. It reveals that there is a greater improvement potential for text embedding space, as our design has no particular preference for the language modality. These findings demonstrate the benefits of our proposed loss for enhancing the joint embedding space across modalities.

5.3. Comparison on Continue Pre-training

We build upon OpenAI’s publicly available CLIP¹ as the baseline and conduct continuous pre-training using our proposed loss on the LLaVA-595K dataset [41]. The superiority of our approach is demonstrated in Table 3. Notably, the most substantial improvement is observed in the image-text retrieval task, with a Recall@1 increase of approximately +3% to +4%. This improvement is approximately three

¹<https://github.com/openai/CLIP>

times larger than the improvement observed in the classification task. Here, the significant enhancement is actually due to the fact that the retrieval task places a higher demand on the quality of the affinity. Our proposed approach better aligns the affinities between images and their corresponding text descriptions, resulting in better affinity affinities and improved retrieval performance.

5.4. Ablation Study

Component Effect. Table 4 shows that both SaCo loss and pseudo-affinity label mimicking are essential components for achieving the best performance in our approach. Mimicking the pseudo affinity alone only brings a minor improvement, as it does not effectively constrain the cross-modal consistency of sample-wise affinity. This highlights the importance of cross-modal affinity consistency as an indispensable objective, while pseudo-affinity mimicking contributes to training stability and convergence to a better local optimum. For continue pre-training, solely using our

Pre-train Schedule	SaCo Loss	Pseudo Affinity	ImageNet-1K		Flickr30K	
			Top-1 Acc.	I2T(R@1)	T2I(R@1)	I2T(R@1)
From Scratch	✓	✓	11.9	16.0	12.2	
			17.5 (+5.6)	23.2 (+7.2)	16.3 (+4.1)	
	✓	✓	12.4 (+0.5)	17.8 (+1.8)	13.7 (+1.5)	
	✓	✓	18.3 (+6.4)	25.3 (+9.3)	18.3 (+6.1)	
Continue	✓	✓	63.0	77.8	57.0	
			64.0 (+1.0)	81.7 (+3.9)	64.9 (+7.9)	
	✓	✓	60.8 (-2.2)	76.3 (-1.5)	52.6 (-4.4)	
	✓	✓	63.2 (+0.2)	78.6 (+0.8)	59.8 (+2.8)	

Table 4. Ablation on the effect of each component. The CLIP with ViT-B32 image encoder is used as the baseline model. We report the Top-1 Accuracy (Acc.) for zero-shot classification, and Recall@1 (R@1) for both zero-shot image-to-text (I2T) and text-to-image (T2I) retrieval.

SaCo loss performs best while mimicking pseudo-affinity leads to a significant decline in overall performance. This can be attributed to the fact that a well-trained model has already learned a relatively optimal sample-wise affinity, and forcibly aligning its affinity with the pseudo-affinity predicted by another model will excessively disrupt its embedding space.

Investigate the Pseudo-affinity. Table 5 highlights several important findings. Firstly, replicating pseudo-affinity in both modalities simultaneously (row 1 vs 7) conflicts with the purpose of our approach, which is to address the poor cross-modal consistency of pseudo-affinity in existing models. Secondly, mimicking visual pseudo-affinity is better than linguistic counterpart due to noisy textual information in image-text datasets (see row 2-3 vs 7). Images provide more comprehensive information. Thirdly, vision-language models like CLIP [55] yield better pseudo-affinity than vision-only models like DINOv2 [49]. ImageNet-1K pre-trained models have the worst results (see line 9,10) due to limited semantic concepts in the training data. In addition, It is better to use a similar architecture vision model to generate pseudo-affinity for training vision-language models.

Loss Weights. Figure 6 shows the effect of weight coefficient α and β corresponding to our SaCo loss and pseudo-affinity mimicking, respectively. For both of them, the performance shows an unimodal shape as weight varies. We empirically set $\alpha = \beta = 5$.

6. Conclusion

This paper investigates and analyzes a problem of inconsistent cross-modal affinity in vision-language models, namely “Affinity Inconsistency Problem”. To solve this problem, we propose a novel loss function, named Sample-wise affinity Consistency (SaCo) loss, which aims to enhance the

	Model	Pseudo Affinity Source		ImageNet-1K		Flickr30K	
		Modality	Model	Top-1 Acc.	I2T R@1	T2I R@1	
1	CLIP ViT-B/16	Vision-Language	CLIP ViT-L/14	18.5	26.1	20.1	
2		Language	CLIP-Text	20.7	32.5	24.5	
3			BERT-Base	20.5	31.5	23.3	
4	CLIP ViT-B/16	Vision	R50 [†]	20.6	32.4	23.7	
5			ViT-B/32 [†]	21.3	33.6	25.8	
6			ViT-B/16 [†]	21.7	34.0	26.2	
7			ViT-L/14 [†]	21.8	34.7	26.7	
8			ViT-L/14 [*]	21.6	34.0	26.3	
9			ViT-L/14 [‡]	19.4	29.8	22.0	
10	CLIP R50	Vision	R50 [‡]	18.8	31.0	26.0	
11			R50 [†]	22.5	42.0	31.8	
12			ViT-B/16 [†]	22.5	40.1	30.8	
13			ViT-L/14 [†]	22.2	39.5	30.7	

Table 5. Ablation on the effect of pseudo-affinity. All the experiments are pre-trained from scratch on CC3M. We report the Top-1 Accuracy (Acc.) for zero-shot classification, and Recall@1 (R@1) for both zero-shot image-to-text (I2T) and text-to-image (T2I) retrieval. “[†]” denotes the model is supervised pre-trained on ImageNet-1K. “[‡]” represents the official CLIP [55] published by OpenAI. “^{*}” indicates the model is self-supervised pre-trained via DINOv2 [49].

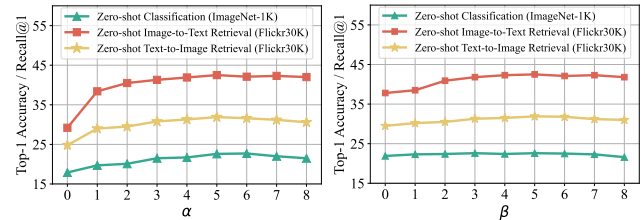


Figure 6. Ablation on loss weights. α is the weight coefficient for our SaCo loss. β corresponds to the pseudo-affinity mimicking objective.

sample-wise affinity consistency by minimizing the distance between image embedding similarity and text embedding similarity for any two samples. Our loss can either be used to pre-train models from scratch or applied to well-trained models via continue pre-training. Extensive experiments indicate that our SaCo loss can bring significant improvement in a broad range of tasks, which highlights the importance of sample-wise affinity consistency across different modalities.

Acknowledgements. This work was supported in part by the Research Grants Council under the Areas of Excellence scheme grant AoE/E-601/22-R, Hong Kong Research Grant Council - Early Career Scheme (Grant No. 27209621), General Research Fund Scheme (Grant No. 17202422), and RGC Matching Fund Scheme (RMGS).

References

- [1] Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, Songhao Piao, and Furu Wei. VLMo: Unified vision-language pre-training with mixture-of-modality-experts. In *Advances in Neural Information Processing Systems*, 2022. 2
- [2] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3558–3568, 2021. 2
- [3] Ying-Cong Chen, Xiaogang Xu, Zhuotao Tian, and Jiaya Jia. Homomorphic latent space interpolation for unpaired image-to-image translation. In *CVPR*, 2019. 2
- [4] 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. 2
- [5] Jifeng Dai, Min Shi, Weiyun Wang, Sitong Wu, Linjie Xing, Wenhui Wang, Xizhou Zhu, Lewei Lu, Jie Zhou, Xiaogang Wang, et al. Demystify transformers & convolutions in modern image deep networks. *arXiv preprint arXiv:2211.05781*, 2022. 2
- [6] Zhiyuan Dang, Cheng Deng, Xu Yang, Kun Wei, and Heng Huang. Nearest neighbor matching for deep clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 3
- [7] Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tiejian Liu, and Wei-Ying Ma. Dual learning for machine translation. In *Advances in Neural Information Processing Systems*, 2016. 2
- [8] Jianfeng Dong, Yabing Wang, Xianke Chen, Xiaoye Qu, Xirong Li, Yuan He, and Xun Wang. Reading-strategy inspired visual representation learning for text-to-video retrieval. *IEEE transactions on circuits and systems for video technology*, 32(8):5680–5694, 2022. 3
- [9] Xiaoyi Dong, Jianmin Bao, Yinglin Zheng, Ting Zhang, Dongdong Chen, Hao Yang, Ming Zeng, Weiming Zhang, Lu Yuan, Dong Chen, et al. Maskclip: Masked self-distillation advances contrastive language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10995–11005, 2023. 1, 2
- [10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2, 6
- [11] Lijie Fan, Dilip Krishnan, Phillip Isola, Dina Katabi, and Yonglong Tian. Improving clip training with language rewrites. *arXiv preprint arXiv:2305.20088*, 2023. 1
- [12] Haotian Zhang, Long Mai, Ning Xu, Zhaowen Wang, John Collomosse, and Hailin Jin. An internal learning approach to video inpainting. In *ICCV*, 2019. 3
- [13] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 2
- [14] Runhui Huang, Yanxin Long, Jianhua Han, Hang Xu, Xiwen Liang, Chunjing Xu, and Xiaodan Liang. Nlip: Noise-robust language-image pre-training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 926–934, 2023. 1
- [15] Ahmet Iscen, Jack Valmadre, Anurag Arnab, and Cordelia Schmid. Learning with neighbor consistency for noisy labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 3
- [16] 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. 3
- [17] Li Jiang, Shaoshuai Shi, Zhuotao Tian, Xin Lai, Shu Liu, Chi-Wing Fu, and Jiaya Jia. Guided point contrastive learning for semi-supervised point cloud semantic segmentation. In *ICCV*, 2021. 2
- [18] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle consistent adversarial networks. In *ICCV*, 2017. 2
- [19] Xin Lai, Zhuotao Tian, Li Jiang, Shu Liu, Hengshuang Zhao, Liwei Wang, and Jiaya Jia. Semi-supervised semantic segmentation with directional context-aware consistency. In *CVPR*, 2021. 3
- [20] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. LISA: reasoning segmentation via large language model. *arXiv preprint arXiv:2308.00692*, 2023. 2
- [21] Zhengfeng Lai, Haotian Zhang, Wentao Wu, Haoping Bai, Aleksei Timofeev, Xianzhi Du, Zhe Gan, Jiulong Shan, Chen-Nee Chuah, Yinfei Yang, et al. From scarcity to efficiency: Improving clip training via visual-enriched captions. *arXiv preprint arXiv:2310.07699*, 2023. 1
- [22] Chenyang Lei, Yazhou Xing, and Qifeng Chen. Blind video temporal consistency via deep video prior. In *Advances in Neural Information Processing Systems*, 2020. 3
- [23] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. Filip: Fine-grained interactive language-image pre-training. In *ICLR*, 2022. 1, 2
- [24] Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *Proceedings of the AAAI conference on artificial intelligence*, pages 11336–11344, 2020. 2
- [25] Juncheng Li, Xin He, Longhui Wei, Long Qian, Linchao Zhu, Lingxi Xie, Yueting Zhuang, Qi Tian, and Siliang Tang. Fine-grained semantically aligned vision-language pre-training. *Advances in neural information processing systems*, 35:7290–7303, 2022. 2

- [26] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022. 1, 2, 3, 4
- [27] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 1, 2, 3
- [28] 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. 2
- [29] Tianhong Li and Sangnie Bhardwaj. Leveraging unpaired data for vision-language generative models via cycle consistency, 2023. 2
- [30] 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. 2
- [31] 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 *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, pages 121–137. Springer, 2020. 2
- [32] Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. *arXiv preprint arXiv:2110.05208*, 2021. 1, 2, 6
- [33] Yanghao Li, Haoqi Fan, Ronghang Hu, Christoph Feichtenhofer, and Kaiming He. Scaling language-image pre-training via masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 1, 2
- [34] Fangjian Lin, Sitong Wu, Yizhe Ma, and Shengwei Tian. Full-scale selective transformer for semantic segmentation. In *Proceedings of the Asian Conference on Computer Vision*, pages 2663–2679, 2022. 3
- [35] Fangjian Lin, Zhanhao Liang, Sitong Wu, Junjun He, Kai Chen, and Shengwei Tian. Structtoken: Rethinking semantic segmentation with structural prior. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023. 3
- [36] Fangjian Lin, Yizhe Ma, Sitong Wu, Long Yu, and Shengwei Tian. Axwin transformer: A context-aware vision transformer backbone with axial windows. *arXiv preprint arXiv:2305.01280*, 2023. 2
- [37] Fangjian Lin, Jianlong Yuan, Sitong Wu, Fan Wang, and Zhibin Wang. Uninext: Exploring a unified architecture for vision recognition. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 3200–3208, 2023. 2
- [38] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 2, 6
- [39] Baolong Liu, Qi Zheng, Yabing Wang, Minsong Zhang, Jianfeng Dong, and Xun Wang. Featinter: exploring fine-grained object features for video-text retrieval. *Neurocomputing*, 496:178–191, 2022. 3
- [40] Chang Liu, Shaofeng Zhang, Xiaokang Yang, and Junchi Yan. Self-supervised learning of visual graph matching. In *ECCV*, 2022. 3
- [41] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023. 7
- [42] Jiaming Liu, Senqiao Yang, Peidong Jia, Ming Lu, Yandong Guo, Wei Xue, and Shanghang Zhang. Vida: Homeostatic visual domain adapter for continual test time adaptation. *arXiv preprint arXiv:2306.04344*, 2023. 3
- [43] Xiaoliu Luo, Zhuotao Tian, Taiping Zhang, Bei Yu, Yuan Yan Tang, and Jiaya Jia. Pfenet++: Boosting few-shot semantic segmentation with the noise-filtered context-aware prior mask. *TPAMI*, 2024. 3
- [44] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason J. Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and vqa. *arXiv preprint arXiv:1909.11059*, 2019. 2
- [45] Yizhe Ma, Fangjian Lin, Sitong Wu, Shengwei Tian, and Long Yu. Prseg: A lightweight patch rotate mlp decoder for semantic segmentation. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023. 3
- [46] Manuel Lang, Oliver Wang, Tunc Aydin, Aljoscha Smolic, and Markus Gross. Practical temporal consistency for image-based graphics applications. *ACM Transactions on Graphics*, 2012. 3
- [47] Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. Slip: Self-supervision meets language-image pre-training. In *European Conference on Computer Vision*, pages 529–544. Springer, 2022. 2
- [48] Nicolas Bonneel, James Tompkin, Kalyan Sunkavalli, Deqing Sun, and Sylvain Paris. Blind video temporal consistency. *ACM Transactions on Graphics*, 2015. 3
- [49] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 5, 8
- [50] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 6
- [51] Bohao Peng, Zhuotao Tian, Xiaoyang Wu, Chengyao Wang, Shu Liu, Jingyong Su, and Jiaya Jia. Hierarchical dense correlation distillation for few-shot segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23641–23651, 2023. 3

- [52] 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. 2, 6
- [53] Shraman Pramanick, Li Jing, Sayan Nag, Jiachen Zhu, Hardik Shah, Yann LeCun, and Rama Chellappa. Volta: Vision-language transformer with weakly-supervised local-feature alignment. *arXiv preprint arXiv:2210.04135*, 2022. 1, 2
- [54] Di Qi, Lin Su, Jia Song, Edward Cui, Taroon Bharti, and Arun Sacheti. Imagebert: Cross-modal pre-training with large-scale weak-supervised image-text data. *arXiv preprint arXiv:2001.07966*, 2020. 3
- [55] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 2, 3, 4, 5, 6, 8
- [56] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018. 3, 6
- [57] Zhao Shizhen, Gao Changxin, Shao Yuanjie, Li Lerenhan, Yu Changqian, Ji Zhong, and Sang Nong. Gtnet: Generative transfer network for zero-shot object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2020. 6
- [58] Haoru Tan, Sitong Wu, and Jimin Pi. Semantic diffusion network for semantic segmentation. *Advances in Neural Information Processing Systems*, 35:8702–8716, 2022. 3
- [59] Haoru Tan, Chuang Wang, Sitong Wu, Xu-Yao Zhang, Fei Yin, and Cheng-Lin Liu. Ensemble quadratic assignment network for graph matching. *arXiv preprint arXiv:2403.06457*, 2024. 3
- [60] Haoru Tan, Sitong Wu, Fei Du, Yukang Chen, Zhibin Wang, Fan Wang, and Xiaojuan Qi. Data pruning via moving-one-sample-out. *Advances in Neural Information Processing Systems*, 36, 2024. 2
- [61] Hao-Ru Tan, Chuang Wang, Si-Tong Wu, Tie-Qiang Wang, Xu-Yao Zhang, and Cheng-Lin Liu. Proxy graph matching with proximal matching networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9808–9815, 2021. 3
- [62] Zhuotao Tian, Michelle Shu, Pengyuan Lyu, Ruiyu Li, Chao Zhou, Xiaoyong Shen, and Jiaya Jia. Learning shape-aware embedding for scene text detection. In *CVPR*, 2019. 3
- [63] Zhuotao Tian, Xin Lai, Li Jiang, Shu Liu, Michelle Shu, Hengshuang Zhao, and Jiaya Jia. Generalized few-shot semantic segmentation. In *CVPR*, 2022.
- [64] Zhuotao Tian, Hengshuang Zhao, Michelle Shu, Zhicheng Yang, Ruiyu Li, and Jiaya Jia. Prior guided feature enrichment network for few-shot segmentation. *TPAMI*, 2022.
- [65] Zhuotao Tian, Jiequan Cui, Li Jiang, Xiaojuan Qi, Xin Lai, Yixin Chen, Shu Liu, and Jiaya Jia. Learning context-aware classifier for semantic segmentation. In *AAAI*, 2023. 3
- [66] Thomas Verelst, Paul K. Rubenstein, Marcin Eichner, Tinne Tuytelaars, and Maxim Berman. Spatial consistency loss for training multi-label classifiers from single-label annotations. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3879–3889, 2023. 3
- [67] Alex Jinpeng Wang, Kevin Qinghong Lin, David Junhao Zhang, Stan Weixian Lei, and Mike Zheng Shou. Too large; data reduction for vision-language pre-training. *arXiv preprint arXiv:2305.20087*, 2023. 1
- [68] Luozhou Wang, Shuai Yang, Shu Liu, and Ying cong Chen. Not all steps are created equal: Selective diffusion distillation for image manipulation, 2023. 2
- [69] Yabing Wang, Jianfeng Dong, Tianxiang Liang, Minsong Zhang, Rui Cai, and Xun Wang. Cross-lingual cross-modal retrieval with noise-robust learning. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 422–433, 2022. 3
- [70] Yabing Wang, Fan Wang, Jianfeng Dong, and Hao Luo. Cl2cm: Improving cross-lingual cross-modal retrieval via cross-lingual knowledge transfer. *arXiv preprint arXiv:2312.08984*, 2023.
- [71] Yabing Wang, Shuhui Wang, Hao Luo, Jianfeng Dong, Fan Wang, Meng Han, Xun Wang, and Meng Wang. Dual-view curricular optimal transport for cross-lingual cross-modal retrieval. *IEEE Transactions on Image Processing*, 33:1522–1533, 2024. 3
- [72] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*, 2021. 3
- [73] Lin Wu, Yang Wang, and Ling Shao. Cycle-consistent deep generative hashing for cross-modal retrieval. *IEEE Transactions on Image Processing*, 2019. 3
- [74] Sitong Wu, Tianyi Wu, Fangjian Lin, Shengwei Tian, and Guodong Guo. Fully transformer networks for semantic image segmentation. *arXiv preprint arXiv:2106.04108*, 2021. 3
- [75] Sitong Wu, Tianyi Wu, Haoru Tan, and Guodong Guo. Pale transformer: A general vision transformer backbone with pale-shaped attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2731–2739, 2022. 2
- [76] Xiaoyang Wu, Zhuotao Tian, Xin Wen, Bohao Peng, Xihui Liu, Kaicheng Yu, and Hengshuang Zhao. Towards large-scale 3d representation learning with multi-dataset point prompt training. *arXiv preprint arXiv:2308.09718*, 2023. 3
- [77] Xuan Dong, Boyan Bonev, Yu Zhu, and Alan L Yuille. Region-based temporally consistent video post-processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2015. 3
- [78] Xuyang Bai, Zixin Luo, Lei Zhou, Hongkai Chen, Lei Li, Zeyu Hu, Hongbo Fu, and Chiew-Lan Tai. Pointdsc: Robust point cloud registration using deep spatial consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 3

- [79] Senqiao Yang, Jiaming Liu, Ray Zhang, Mingjie Pan, Zoey Guo, Xiaoqi Li, Zehui Chen, Peng Gao, Yandong Guo, and Shanghang Zhang. Lidar-llm: Exploring the potential of large language models for 3d lidar understanding. *arXiv preprint arXiv:2312.14074*, 2023. 3
- [80] Senqiao Yang, Tianyuan Qu, Xin Lai, Zhuotao Tian, Bohao Peng, Shu Liu, and Jiaya Jia. An improved baseline for reasoning segmentation with large language model. *arXiv preprint arXiv:2312.17240*, 2023. 2
- [81] Senqiao Yang, Jiarui Wu, Jiaming Liu, Xiaoqi Li, Qizhe Zhang, Mingjie Pan, and Shanghang Zhang. Exploring sparse visual prompt for cross-domain semantic segmentation. *arXiv preprint arXiv:2303.09792*, 2023. 3
- [82] Shuai Yang, Yukang Chen, Luozhou Wang, Shu Liu, and Yingcong Chen. Denoising diffusion step-aware models, 2024. 2
- [83] Yuting Gao, Jinfeng Liu, Zihan Xu, Jun Zhang, Ke Li, Rongrong Ji, and Chunhua Shen. Pyramidclip: Hierarchical feature alignment for vision-language model pretraining. In *Advances in Neural Information Processing Systems*, 2022. 2
- [84] Yuting Gao, Jinfeng Liu, Zihan Xu, Tong Wu, Wei Liu, Jie Yang, Ke Li, and Xing Sun. Softclip: Softer cross-modal alignment makes clip stronger. *arXiv preprint arXiv:2303.17561*, 2023. 1, 2
- [85] Zelin Zang, Lei Shang, Senqiao Yang, Fei Wang, Baigui Sun, Xuansong Xie, and Stan Z Li. Boosting novel category discovery over domains with soft contrastive learning and all in one classifier. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11858–11867, 2023. 2
- [86] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18123–18133, 2022. 3
- [87] Shaofeng Zhang, Qiang Zhou, Zhibin Wang, Fan Wang, and Junchi Yan. Patch-level contrastive learning via positional query for visual pre-training. In *ICML*, 2023. 2
- [88] Shaofeng Zhang, Feng Zhu, Rui Zhao, and Junchi Yan. Contextual image masking modeling via synergized contrasting without view augmentation for faster and better visual pre-training. In *ICLR*, 2023. 2
- [89] Shaofeng Zhang, Feng Zhu, Rui Zhao, and Junchi Yan. Patch-level contrasting without patch correspondence for accurate and dense contrastive representation learning. *ICLR*, 2023. 2
- [90] Shaofeng Zhang, Jinfeng Huang, Qiang Zhou, Zhibin Wang, Fan Wang, Jiebo Luo, and Junchi Yan. Continuous-multiple image outpainting in one-step via positional query and a diffusion-based approach. 2024. 2
- [91] Qi Zheng, Jianfeng Dong, Xiaoye Qu, Xun Yang, Yabing Wang, Pan Zhou, Baolong Liu, and Xun Wang. Progressive localization networks for language-based moment localization. *ACM Transactions on Multimedia Computing, Communications and Applications*, 19(2):1–21, 2023. 3
- [92] Qiang Zhou, Chaohui Yu, Shaofeng Zhang, Sitong Wu, Zhibing Wang, and Fan Wang. Regionblip: A unified multi-modal pre-training framework for holistic and regional comprehension. *arXiv preprint arXiv:2308.02299*, 2023. 2