TiMix: Text-Aware Image Mixing for Effective Vision-Language Pre-training

Chaoya Jiang[1], Wei Ye[1*], Haiyang Xu[2*], Qinghao Ye[2], Ming Yan[2], Ji Zhang[2], Shikun Zhang[1]

[1]National Engineering Research Center for Software Engineering, Peking University, Beijing, China
[2]Alibaba Group, Hangzhou, China
wye@pku.edu.cn, shuofeng.xhy@alibaba-inc.com

Abstract

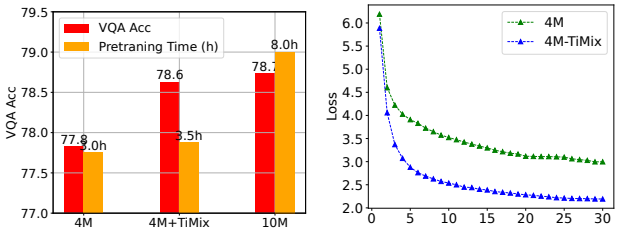
Self-supervised Multi-modal Contrastive Learning (SMCL) remarkably advances modern Vision-Language Pre-training (VLP) models by aligning visual and linguistic modalities. Due to noises in web-harvested text-image pairs, however, scaling up training data volume in SMCL presents considerable obstacles in terms of computational cost and data inefficiency. To improve data efficiency in VLP, we propose Text-aware Image Mixing (TiMix), which integrates mix-based data augmentation techniques into SMCL, yielding significant performance improvements without significantly increasing computational overhead. We provide a theoretical analysis of TiMix from a mutual information (MI) perspective, showing that mixed data samples for cross-modal contrastive learning implicitly serve as a regularizer for the contrastive loss. The experimental results demonstrate that TiMix exhibits a comparable performance on downstream tasks, even with a reduced amount of training data and shorter training time, when benchmarked against existing methods. This work empirically and theoretically demonstrates the potential of data mixing for data-efficient and computationally viable VLP, benefiting broader VLP model adoption in practical scenarios. Our code is available on https://github.com/chaoyajiang/TiMiX/tree/main.

Introduction

Vision-Language Pre-training (VLP) exploits large-scale image-text pairs without annotations via self-supervised learning (Syed, Gaol, and Matsuo 2021; Liu et al. 2020), achieving tremendous success on a wide range of cross-modal downstream tasks (Chen et al. 2020; Huang et al. 2020; Li et al. 2020; Yu et al. 2021; Li et al. 2021; Zhang et al. 2021; Kim, Son, and Kim 2021; Li et al. 2022a; Xu et al. 2021; **?**). More recently, Self-supervised Multi-modal Contrastive Learning (SMCL) has emerged as a significant advancement in the VLP community (Li et al. 2022b; Radford et al. 2021b; Li et al. 2021, 2022a; Zeng, Zhang, and Li 2021), facilitating the learning of cross-modal representations from image-text pairs by aligning visual and linguistic modalities.

Recent studies (Li et al. 2021, 2022b; Jiang et al. 2023c) have found that SMCL-based models pre-trained on web-harvested data often suffer from data inefficiency since image captions frequently contain words that are unrelated to the image content or only capture partial information. One common strategy is increasing the scale of the training to alleviate the negative impacts of noisy data samples (Radford et al. 2021b; Jia et al. 2021). A typical example is CLIP (Radford et al. 2021b), which utilizes a massive dataset of 400 million image-text pairs obtained through web crawling. Though it demonstrated promising results in enhancing the cross-modal capabilities of models, scaling up datasets presents a challenge due to the high computational cost. For example, CLIP requires an estimated 3584 GPU (V100) days for pertaining, a demand that is financially prohibitive under a constrained budget. Other researchers exploit soft labels (Li et al. 2021) or regenerate image captions (Li et al. 2022b) to mitigate the impact of noisy data, yet with unsatisfactory performance improvement or substantial additional computation.

In this paper, we present a novel perspective of data mixing to tackle data inefficiency in VLP. We hypothesize that an image could exhibit multiple distinct views, each potentially associated with a different textual caption. These diverse textual descriptions align with specific views that capture various aspects of the image's semantic information. Building upon this hypothesis, we introduce Text-aware Image Mixing



(a) VQA and Pre-training Time (b) Contrastive Learning Loss

Figure 1: Subfigure (a) illustrates the Visual Question Answering results and pre-training time per epoch of the VLP model mPLUG (Li et al. 2022a) which are pre-trained on with different data sizes on 8 × 80G A100. 4M+TiMix refers to training on 4M data with TiMix. Subfigure (b) illustrates the convergence curve of cross-modal contrastive learning, the x-axis is labeled as epoch.

* Corresponding Author.

(TiMix), which adopts the CutMix (Yun et al. 2019) approach to create data samples for contrastive learning. Specifically, we design a patch-text alignment (PTA) pre-training task, allowing us to learn the matching degree between patches and captions. So we can mix two images guided by the relevance of their patches to their captions. Then the mixed samples are incorporated into contrastive learning to improve cross-modal representation and enhance data efficiency.

We theoretically analyze TiMix from a mutual information (MI) maximation perspective and find that mixed data samples implicitly provide a regularizer for the contrastive learning loss function. This regularizer keeps the model from overfitting to partially aligned image-text pairs during the contrastive learning process, thereby mitigating the negative impact of noisy data. Empirically, by incorporating TiMix into existing VLP models, we can observe consistent performance improvement on common vision-language downstream tasks, including Visual Question Answering (VQA), Cross-modal Retrieval, Natural Language for Visual Reasoning (NLVR) and Image Captioning, with small additional computational cost during training.

In summary, our contributions are:

- We take the first step to introduce mix-based data samples into vision-language pre-training. With a novel patch-text alignment pre-training task, mixed images are created in a CutMix style based on the matching degree of their patches and captions, serving as high-quality data for cross-modal contrastive learning.
- We theoretically prove that mixed data samples implicitly provide a regularizer for cross-modal contrastive learning, facilitating mutual information optimization for potentially partially-aligned image-text pairs.
- Experimental findings illustrate that TiMix delivers robust performance, significantly enhancing data efficiency while maintaining cost-effectiveness during the pre-training phase. For example, as shown in Figure 1, TiMix achieves comparable downstream task performance by training on 40% of the data in 43.8% of the training time, compared to a recent robust VLP model mPLUG.

Related Work

Vision-Language pre-training

Recent years have seen significant success for large-scale pre-trained vision-language models (Tan and Bansal 2019; Chen et al. 2020; Huang et al. 2020; Li et al. 2020; Yu et al. 2021; Li et al. 2021; Wang et al. 2021b; Li et al. 2022a; Zhang et al. 2021; Jiang et al. 2023a,b) in a variety of cross-modal tasks. Current approaches to VLP can be broadly divided into two categories in terms of visual representation extraction. The first category is detector-based VLP methods (Li et al. 2019; Tan and Bansal 2019; Li et al. 2020; Chen et al. 2020; Yu et al. 2021; Fang et al. 2021). These methods primarily adopt a two-step training pipeline: they extract visual features using a pre-trained object detector and then train the cross-modal pre-training model to align text and visual features. The main challenge for these methods is to balance effectiveness and efficiency. The second category consists of more recent CNN-based (Xu et al. 2021) or ViTs-based (Li et al. 2021; Kim, Son, and Kim 2021; Radford et al. 2021a) methods, especially patch-based ViT. These methods eliminate the need for a complex object detector in feature extraction, enabling end-to-end VL learning. Furthermore, Self-supervised Multi-modal Contrastive Learning (SMCL) has lately sparked significant advancements (Li et al. 2022b; Radford et al. 2021b; Yao et al. 2021; Li et al. 2021, 2022a) by conducting cross-modal alignment. SMCL consists of image-to-text and text-to-image contrastive learning, e.g., with the InfoNCE (Oord, Li, and Vinyals 2018) loss.

Mixed Data Augmentation

Mixup (Zhang et al. 2017) is a widely used data augmentation technique in Computer Vision, which involves training by convexly combining the input image and its corresponding label. CutMix (Yun et al. 2019), on the other hand, is a specific case of Mixup and can be seen as a pixel-level Mixup method that utilizes binary masks. Recent developments (Uddin et al. 2020; Liu et al. 2022; Walawalkar et al. 2020) in Mixup methods have focused on effectively leveraging saliency information and performing mixing at the image feature level. The motivation behind saliency-based Mixup methods is to preserve salient regions when blending images, ensuring that sufficient information is retained and more natural feature representations are learned. SaliencyMix (Uddin et al. 2020) employs various saliency detectors to directly extract salient regions from the images. Co-Mixup (Kim et al. 2021) aims to maximize the gradient-based saliency while encouraging diversity in the mixed images' hypermodules. SuperMix (Dabouei et al. 2020) utilizes supervised signals to mix input images based on saliency. However, the majority of existing Mixup methods have mainly been applied in the CV. In contrast, we introduce Mixup techniques to the multimodal domain for the first time, resulting in notable advancements.

Method

Our approach exhibits similarities to the Cutmix-style method. Specifically, our method involves pairing two images within a batch, denoted as I^x and I^y, along with their corresponding texts, T^x and T^y. Assuming I^y as the source image, we extract the region $\dot{R}^y$ in the source image I^y that exhibits the highest text-relevant score with T^y. Subsequently, we replace the corresponding region $\dot{R}^x$ in the target image I^x, which possesses the lowest text-relevant score with T^x, with the extracted region $\dot{R}^y$. This process yields a mixed image denoted as $\hat{I}^{x,y}$. To determine the soft labels for the mixed image and text, we consider the proportion of the mixed and cropped regions. Additional details regarding this process are provided in the following section.

Text-aware Region Mixing

As depicted in Figure 2, considering two pairs of image-text denoted as I^x, T^x and I^y, T^y respectively, let us designate I^x as the target image and I^y as the source image. Initially, suppose the image size is $H \times W$, the target image I^x and the source image I^y are divided into distinct, non-overlapping patches of size $P \times P$, resulting in a total of $\frac{H}{P} \times \frac{W}{P}$ patches. Subsequently, these patches are fed into

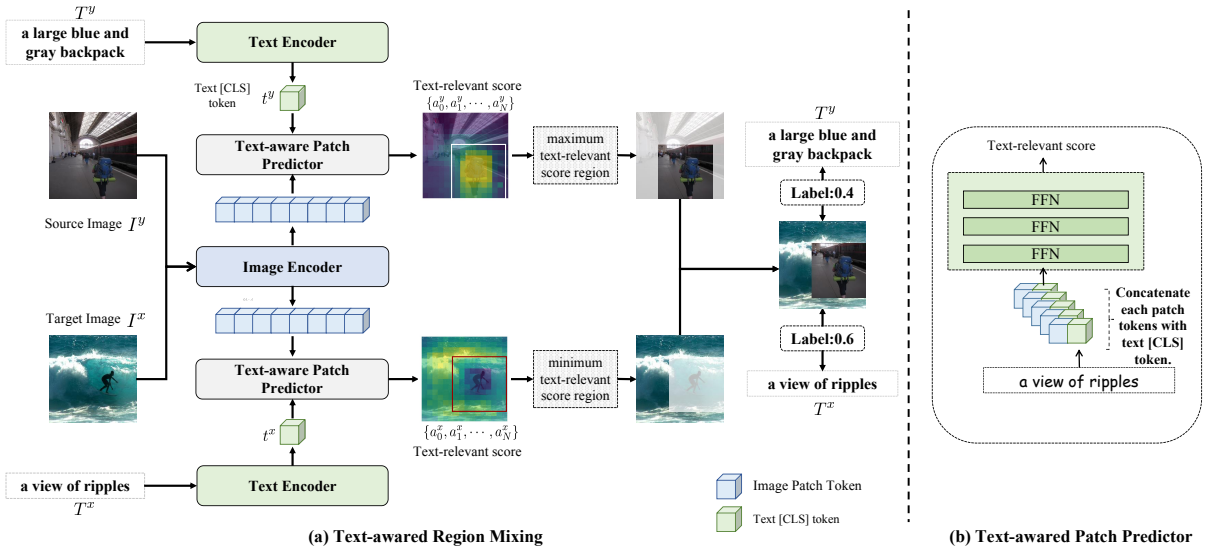


Figure 2: The subfigure (a) illustrates the process of TiMix, where two image-text pairs are utilized. Subfigure (b) depicts the architecture of the Text-aware patch predictor.

the visual backbone to extract the sequential representation of visual features. Consequently, the extracted representations of I^x and I^y assume the form of $\{v^x_{cls}, v^x_1, \ldots, v^x_N\}$ and $\{v^y_{cls}, v^y_1, \ldots, v^y_N\}$ respectively, where $N = \frac{H}{P} \times \frac{W}{P}$. Then, both textual captions T^x and T^y are fed to the text backbone and the [CLS] tokens symbolized as $t^x \in R^D$ and $t^y \in R^D$ are employed to encapsulate the comprehensive context of the text, D is the dimension of text representation. Next, each visual token in the sequence $\{v^x_{cls}, v^x_1, \ldots, v^x_N\}$ is concatenated with t^x and passed through the Text-aware Patch Predictor (TPP) to compute the text-relevant scores for the patches, denoted as $A_x = \{a^x_1, \ldots, a^x_N\}$. Similarly, for the sequence $\{v^y_{cls}, v^y_1, \ldots, v^y_N\}$, each visual token is concatenated with t^y, and fed to the text-relevant scores to calculate the text-relevant scores of patches $A_y = \{a^y_1, \ldots, a^y_N\}$, where $A_x, A_y \in R^N$. We rearrange the shape of A_x, A_y to $\frac{H}{P} \times \frac{W}{P}$.

We aim to extract a region $\hat{R}^y$ from the source image I^y and merge it with the target image I^x to create a mixed image $\hat{I}^{x,y}$. For this purpose, we introduce a side ratio γ, which is sampled from a uniform distribution ranging from 0.25 to 0.75. This ratio helps determine the total number of patches, which is given by $\lfloor \gamma \frac{H}{P} \rfloor \times \lfloor \gamma \frac{W}{P} \rfloor$, that will be obtained from the cropped region. Given a paired text, to calculate the overall text-relevant score of each region with the text, We employ a 2D convolution operation on A_x and A_y with a kernel size of $\lfloor \gamma \frac{H}{P} \rfloor \times \lfloor \gamma \frac{W}{P} \rfloor$ and a stride of 1 to iterate through all the regions. The center indices (denoted as a, b) of the regions should satisfy the following conditions:

$$a^x, b^x = \underset{a,b}{\operatorname{argmin}} \sum_{p,q} A^x_{a+p-\lfloor \frac{h}{2} \rfloor, b+q-\lfloor \frac{w}{2} \rfloor} \tag{1}$$

$$a^y, b^y = \underset{a,b}{\operatorname{argmax}} \sum_{p,q} A^y_{a+p-\lfloor \frac{h}{2} \rfloor, b+q-\lfloor \frac{w}{2} \rfloor} \tag{2}$$

Where $h = \lfloor \gamma \frac{H}{P} \rfloor, w = \lfloor \gamma \frac{W}{P} \rfloor, p \in \{0, 1, \ldots, h-1\}. q \in \{0, 1, \ldots, w-1\}$. Then, we obtain the new mixed image sample $\hat{I}^{x,y}$ as follows:

$$\hat{I}^{x,y} = I^x \tag{3}$$

$$\hat{I}^{x,y}_{a^x+p-\lfloor \frac{h}{2} \rfloor, b^x+q-\lfloor \frac{w}{2} \rfloor} = I^y_{a^y+p-\lfloor \frac{h}{2} \rfloor, b^y+q-\lfloor \frac{w}{2} \rfloor} \tag{4}$$

Then the soft label of the mixed image $\hat{I}^{x,y}$ to text t^y is calculated as follow:

$$\mathbf{s}^y = \frac{hwP^2}{HW} \tag{5}$$

The soft label of the mixed image $\hat{I}^{x,y}$ to text t^x is calculated as $\mathbf{s}^x = 1 - \mathbf{s}^y$.

Learning the Text-relevant Score of Patch

The key component of TiMix is the text-aware patch predictor which needs to predict text-relevant scores between the image patches and input text. As shown in Figure 2, the patch predictor is a Multi-Layer Perceptron (MLP) that contains three linear layers and is used to predict the alignment score between patches and the input text.

As the lack of fine-grained patch-text labels to train the text-aware patch predictor, in this sub-section, we propose to convert object-level signals into patch-level ones and introduce a novel pre-training task named Patch Text Alignment which facilitates the patch predictor training and drives our model to learn the fine-grained patch-text alignment. For object objection and visual grounding datasets like COCO(Lin et al. 2014) and VG(Krishna et al. 2016), the object and region generally be paired with a class label or text description. Therefore, we can transfer every object class label to a text description based on a text template such as "This is a [class label]". Thus, for each (object/region) bounding box in an image, we can construct a text description for it. Then, we transform the bounding box annotations to the patch-level

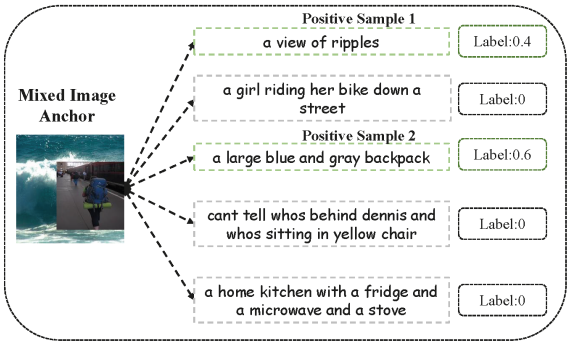


Figure 3: An example of TiMix in image-to-text contrastive learning. The text within the green box represents the positive samples, and the text within the gray box represents the negative samples.

labels by following this rule: Given an image and a bounding box annotation, if there is an overlap between an image patch and a bounding box, it will be assigned with label 1, otherwise, it will be assigned with label 0. For different text descriptions and bounding boxes, the labels of the patch are different. In this way, we can generate fine-grained patch-text labels which can be served as the supervisory signal to pre-train our model.

After that, in each step of pre-training, we randomly sample a mini-batch of images from the object detection/visual grounding datasets. For each image, we randomly select an object/region bounding box and translate the bounding box annotation to the image patch label sequence following the transformation rule we mentioned before. Then, we feed the batch of text descriptions of the bounding boxes and the images together to our VLP model. We hope the text-aware patch predictor can detect all patches which have overlap with the bounding box with the guidance of the bounding box text description. Supposing the text-aware patch predictor has predicted the text-relevant scores between image patches and text, we calculate the binary cross entropy loss between the text-relevant scores and patch labels as:

$$\mathcal{L}_{PTA} = \frac{1}{e}\sum_{i=1}^{e} Y_i log\left(a_i\right) + \left(1 - Y_i\right) log\left(1 - a_i\right) \quad (6)$$

where a_i is the text-relevant score between i_{th} patch in the image and the input text, Y_i is the patch label of i_{th} patch.

Contrastive Learning Based on TiMix

In this subsection, we will introduce how to apply TiMix to Vision Language Pretraining (VLP) with unsupervised cross-modal contrastive learning.

Let us consider a random sampling of N image-text pairs to compose a minibatch. Within this set, we form pairs randomly which results in N groups, each N consisting of two image-text pairs. Let us denote the two pair as I^x, T^x and I^y, T^y. Applying the aforementioned method, we use I^x as the target image and I^y as the source image to generate the mixed image $\hat{I}^{x,y}$. Similarly, by swapping the roles, with I^y as the target image and I^x as the source image, we obtain the mixed image $\hat{I}^{y,x}$. By repeating this process, we acquire N mixed images. As shown in Figure 3, we select a mixed image sample $\hat{I}^{x,y}$ as the anchor. The two texts T^x and T^y associated with the source image and target image are regarded as positive samples. Following the aforementioned procedure, we compute the soft labels of the anchor image to these two positive samples. The remaining texts are considered negative samples. Suppose the extracted global vision representation of the mixed image is denoted as $\hat{v}^{x,y}$ and there are two positive text samples, and the global representations of them are denoted as t^x and t^y. Then, the image-to-text contrastive loss based on TiMix can be formulated as follow:

$$\mathcal{L}^v_{TiMix} = \quad (7)$$
$$-\sum_{i=1:\mathcal{N}}\frac{1}{\mathcal{N}}\left[\mathbf{s}^x\, log\left[\frac{f\left(\hat{v}_i^{x,y}, t_i^x\right)}{f\left(\hat{v}_i^{x,y}, t_i^x\right) + \sum\limits_{t_k \neq t_i^x} f\left(\hat{v}_i^{x,y}, t_k\right)}\right]\right.$$
$$\left. + \mathbf{s}^y\, log\left[\frac{f\left(\hat{v}_i^{x,y}, t_i^y\right)}{f\left(\hat{v}_i^{x,y}, t^y\right) + \sum\limits_{t_k \neq t_i^y} f\left(\hat{v}_i^{x,y}, t_k\right)}\right]\right]$$

where $f\left(\hat{v}^{x,y}, t^x\right)$ measures the distance between $\hat{v}^{x,y}$ and t^x in a semantic space, $\mathcal{N}$ represents the number of batch size. For text-to-image contrastive learning, within a batch, a specific text is paired with its corresponding image, which is used as both the source image and the target image to generate two mixed images. These two mixed images are considered positive samples, while the other mixed samples in the batch are treated as negative samples. The label assigned to the text anchor with respect to these two positive image samples is the same as the labels used in image-to-text contrastive learning.

A Mutual Information Maximization Perspective

In this section, we provide evidence and explanations for our method from the perspective of maximizing mutual information. Following the definition in (Oord, Li, and Vinyals 2018) in the context of image-to-text contrastive learning, the similarity function $f\left(v_i, t_i\right)$ in Equation 8 can be utilized to model the density ratio, which preserves the mutual information between the image v_i and the text t_i and we rewrite the $f\left(v_i, t_i\right)$ to $\frac{P(t_i|v_i)}{P(t_i)}$.

Then, given a batch of unmixed image-text pairs, the vanilla contrastive learning loss $\mathcal{L}^v$ satisfies the following inequality:

$$\mathcal{L}^v = -\underset{t}{E} log\left[\frac{\frac{P(t_i|v_i)}{P(t_i)}}{\frac{P(t_i|v_i)}{P(t_i)} + \sum\limits_{k\neq i}\frac{P(t_k|v_i)}{P(t_k)}}\right] \quad (8)$$
$$\geq -I(t_i, v_i) + log\left(N\right) \quad (9)$$

where $I(t_i, v_i)$ denotes the mutual information between t_i and v_i. The detailed proof can be found in Appendix A. Based on inequality 9, we can get the lower bound of $I(t_i, v_i)$ as:

$$I(t_i, v_i) \geq log\left(N\right) - \mathcal{L}^v \quad (10)$$

With a similar derivation, we can get another inequality about the image-to-text contrastive learning loss in TiMix as fol-

lows:

$$\mathcal{L}^{v}_{TiMix} = -\underset{t}{E}\,\mathbf{s}^{x}\, log\left[\frac{\frac{P(t^{x}|\hat{v}^{x,y})}{P(t^{x})}}{\frac{P(t^{x}|\hat{v}^{x,y})}{P(t^{x})}+\sum_{k\neq i}\frac{P(t_{k}|\hat{v}^{x,y})}{P(t_{k})}}\right]$$
$$-\underset{t}{E}\,\mathbf{s}^{y}\left[\frac{\frac{P(t^{y}|\hat{v}^{x,y})}{P(t^{y})}}{\frac{P(t^{y}|\hat{v}^{x,y})}{P(t^{y})}+\sum_{k\neq j}\frac{P(t_{k}|\hat{v}^{x,y})}{P(t_{k})}}\right] \quad (11)$$
$$\geq -\mathbf{s}^{x}\, I(t^{x},\hat{v}^{x,y}) - \mathbf{s}^{y}\, I(t^{y},\hat{v}^{x,y}) + (\mathbf{s}^{x}+\mathbf{s}^{y}) log\,(N) \quad (12)$$

where the $I(t^x, \hat{v}^{x,y})$ is the mutual information between t^x and $\hat{v}^{x,y}$. Noted that given a mixed image $\hat{v}^{x,y}$, there is a target image v^x paired with the text t^x and a source image v^y paired with the text t^y. Suppose $v^x_{\mathbf{r}}$ denotes the region in v^x that has the maximum text-relevant score with t^x and $v^y_{\mathbf{r}}$ denotes the region in v^y that has the maximum text-relevant score with t^y, the mixed image $\hat{v}^{x,y}$ can be seen as the combination of $v^x_{\mathbf{r}}$ and $v^y_{\mathbf{r}}$. Then, based on the chain rule of mutual information, we can get the lower bound of $\mathbf{s}^x\, I(t^x, v^x_{\mathbf{r}}) + \mathbf{s}^y\, I(t^y, v^y_{\mathbf{r}})$ as:

$$\mathbf{s}^{x}\, I(t^{x},v^{x}_{\mathbf{r}}) + \mathbf{s}^{y}\, I(t^{y},v^{y}_{\mathbf{r}})$$
$$\geq log\,(N) - (\mathcal{L}^{v}_{TiMix} + \mathbf{s}^{x}\, I(t^{x},v^{y}_{\mathbf{r}}) + \mathbf{s}^{y}\, I(t^{y},v^{x}_{\mathbf{r}})) \quad (13)$$

The details of this derivation can be found in appendix A. Combining inequality 10 and 13 provides us with inspirational findings. It has been easily seen (and widely known) that traditional InfoNCE loss tries to maximize the lower bound of mutual information between t_i and v_i. Similarly, in the scenario of TiMix , the lower bound of $I(t^x, v^x_{\mathbf{r}})$ and $I(t^y, v^y_{\mathbf{r}})$ should be maximized given ideal clean data. Note that on the right side of inequality 13, $\mathcal{L}^{v}_{TiMix}$ is attained with two items $I(t^x, v^y_{\mathbf{r}})$ and $I(t^y, v^x_{\mathbf{r}})$. These two elements effectively act as implicit regularizers, preventing $\mathcal{L}^{v}_{TiMix}$ from becoming excessively optimized. This mitigates the risk of over-maximizing the mutual information terms $I(t^x, v^x_{\mathbf{r}})$ and $I(t^y, v^y_{\mathbf{r}})$, thereby making the model more robust in the context of inaccurately or partially aligned image-text pairs and hence leading to an improvement of data efficiency.

Experiment

Following the previous works (Li et al. 2021) and (Li et al. 2022a), we use the same pre-training dataset with 14M images with texts, which includes two in-domain datasets (MS COCO (Lin et al. 2014) and Visual Genome (Krishna et al. 2016)), and three web out-domain datasets (Conceptual Captions (Sharma et al. 2018a), Conceptual 12M (Changpinyo et al. 2021a), SBU Captions (Ordonez, Kulkarni, and Berg 2011)). Please refer to Appendix C to see more detail about the pre-training dataset and pre-training setting.

Overall Performance

We evaluate TiMix with two well-known VLP models ALBEF and mPLUG (denoted as ALBEF-TiMix and mPLUG-TiMix) on four vision-language downstream tasks: visual question answering (VQA2.0 (Agrawal et al. 2015)), natural language for visual reasoning (NLVR2 (Suhr et al. 2018)), image-text retrieval(Flickr30K (Plummer et al. 2015), COCO (Lin et al. 2014)), image captioning(COCO Caption (Lin et al. 2014)). Our baselines cover 16 VLP models, detailed in Appendix E(In our experiments, we only re-implement the base version of ALBEF (Li et al. 2021) and mPLUG (Li et al. 2022a).). We will first analyze their overall performances on these tasks. The fine-tuning hyper-parameters and the details of downstream tasks are described in Appendix D.

model	Pre-train Data	VQA		NLVR2	
		dev	std	Dev	Test-P
VisualBERT	180K	70.80	71.00	67.40	67.00
LXMERT	180K	72.42	72.54	74.90	74.50
ViLBERT	3.3M	70.63	70.92	-	-
E2E-VLP	4M	73.25	73.67	77.25	77.96
UNITER	4M	73.82	74.02	79.12	79.98
METER	4M	77.68	77.64	82.33	83.05
ViLT	4M	71.26	71.29	75.18	76.2
VLMo	4M	76.64	76.89	82.77	83.34
OSCAR	6.5M	73.16	73.44	78.07	78.36
VinVL	5.65M	76.52	76.60	82.67	83.98
XVLM	14M	78.22	78.37	84.41	84.76
BLIP	129M	78.25	78.32	82.48	83.08
MAP	4M	78.03	-	83.30	83.48
SCL	4M	78.72	78.78	83.63	84.27
SimVLM	1.8B	77.87	78.14	81.72	81.77
ALBEF	4M	74.54	74.70	80.24	80.50
ALBEF -TiMix	4M	75.92	76.23	82.42	83.03
ALBEF	14M	75.84	76.04	82.55	83.14
ALBEF -TiMix	14M	76.82	77.11	83.44	83.47
mPLUG	4M	77.83	77.98	82.66	82.92
mPLUG -TiMix	4M	78.63	78.85	84.12	84.23
mPLUG	14M	79.65	79.22	83.43	84.21
mPLUG -TiMix	14M	**80.83**	**81.53**	**84.77**	**85.22**

Table 1: Evaluation results on VQA2.0 and NLVR^2. More details about comparison models are in Appendix E.

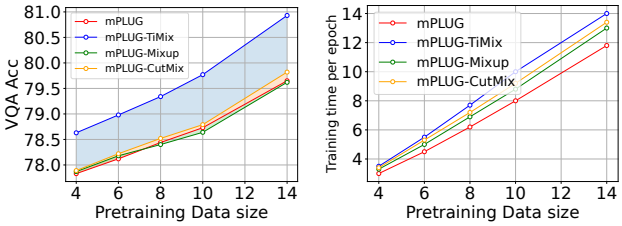
Visual Question Answering The VQA task requires the model to answer natural language questions given an image. Following the approach of (Li et al. 2021), we treat VQA as an answer generation problem. We evaluated our models by submitting the results to the evaluation server [1] and report the test-dev and test-std scores in Table 1. The VLP models equipped with TiMix demonstrate improved performance on the VQA task compared to the models without TiMix . These results highlight the significant improvements achieved by TiMix . Additionally, our mPLUG-TiMix model trained on 14M data outperforms other baseline models which provides further evidence of the effectiveness of our method.

Natural Language for Visual Reasoning The NLVR2 (Suhr et al. 2018) task requires the model to predict whether a sentence accurately describes a pair of images, which is a binary classification task. For ALBEF-TiMix and mPLUG-TiMix , we follow (Li et al. 2021) and use two cross-attention layers to process the two input images; their outputs are merged and fed into a Feed Forward Network (FFN). As shown in Table 1, pre-trained with 14M, mPLUG-TiMix can obtain competitive performances to the SOTA models.

[1] https://eval.ai/web/challenges/challenge-page/830/leaderboard

Models	# Pre-train data	MSCOCO (5K test set)						Flickr30K (1K test set)					
		TR			IR			TR			IR		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
E2E-VLP	4M	-	-	-	-	-	-	86.2	97.5	98.92	73.6	92.4	96.0
OSCAR	4M	70.0	91.1	95.5	54.0	80.8	88.5	-	-	-	-	-	-
VinVL	5.65M	75.4	92.9	96.2	58.8	83.5	90.3	-	-	-	-	-	-
ViLBert	3.3M	-	-	-	-	-	-s	-	-	-	58.2	84.9	91.5
UNITER	4M	65.7	88.6	93.8	52.9	79.9	88.0	87.3	98.0	99.2	75.6	94.1	96.8
METER	4M	76.2	93.2	96.8	57.1	82.7	90.1	94.3	99.6	99.9	82.2	96.3	98.4
VLMo	4M	78.2	94.4	97.4	60.6	84.4	91.0	95.3	99.9	100.0	84.5	97.3	98.6
BLIP	129M	**82.4**	95.4	97.9	65.1	86.3	91.8	97.4	99.8	99.9	87.6	97.7	99.0
SCL	4M	77.7	94.1	97.4	60.1	84.6	91.5	95.9	99.8	100.0	84.6	97.4	98.9
MAP	4M	79.3	94.8	97.6	60.9	86.2	93.1	94.9	99.5	99.8	83.8	97.2	98.7
ViLT	4M	61.5	86.3	92.7	42.7	72.9	83.1	83.5	96.7	98.6	64.4	88.7	93.8
ALBEF	4M	73.1	91.4	96.0	56.8	81.5	89.2	94.3	99.4	99.8	82.8	96.7	98.4
ALBEF -TiMix	4M	76.4	93.7	96.6	60.4	83.2	90.3	95.1	99.8	100.0	84.2	97.3	98.6
ALBEF	14M	77.6	94.3	97.2	60.7	84.3	90.5	95.9	99.8	100.0	85.6	97.5	98.9
ALBEF -TiMix	14M	78.8	95.2	97.6	61.3	85.2	91.0	96.7	99.8	100.0	86.4	97.2	99.0
mPLUG	4M	78.8	94.1	96.4	61.2	85.2	90.6	95.9	99.8	100.0	85.7	96.8	98.6
mPLUG -TiMix	4M	80.5	95.3	97.2	63.3	85.4	91.5	96.8	99.8	100.0	86.2	97.6	98.8
mPLUG	14M	80.6	94.8	97.1	63.9	85.5	91.2	96.5	99.8	100.0	86.3	97.2	98.9
mPLUG -TiMix	14M	82.3	**95.8**	**98.0**	**65.2**	**87.0**	**92.1**	**97.2**	**99.8**	**100.0**	**87.9**	**97.8**	**99.0**

Table 2: Evaluation results of image-text retrieval on Flickr30K (Plummer et al. 2015) and COCO datasets (Lin et al. 2014).



(a) VQA accuracy. (b) Pre-training time per epoch

Figure 4: The visualization of VQA accuracy and Pre-training time per epoch of different models pre-trained on different data sizes

Image-Text Retrieval We conduct experiments for both image-to-text retrieval (TR) and text-to-image retrieval (IR) on MSCOCO (Lin et al. 2014) and Flickr30K (Plummer et al. 2015) datasets. As shown in Table 2, pre-trained with 14M images, mPLUG-TiMix outperforms all existing methods on both datasets which even achieves better performance than BLIP with 129M. In addition, all models equipped with TiMix show significant improvements compared to their counterparts without TiMix.

Image Captioning Following (Li et al. 2022a), we fine-tune mPLUG/ mPLUG-TiMix with cross-entropy loss and then with CIDEr optimization for an additional 5 epochs. Our experiments, as shown in Table 3, unequivocally illustrate the superiority of mPLUG-TiMix over mPLUG alone. Notably, mPLUG-TiMix achieves performance levels that are comparable to those of SOTA models.

Ablation Study

We conducted ablation studies to examine the impact of the Pretraining task Patch Text Alignment (PTA) and mix-based data augmentation. Specifically, we investigated the effects of removing the PTA task while keeping the mix-based data augmentation for contrastive learning (**w/o PTA**). Without the PTA task, the Text-aware Patch Predictor cannot be optimized effectively, so we replaced it with a simple strategy where we follow the way of CutMix(Yun et al. 2019) and randomly sample the region in the image. In Table 4, we can observe that without the text guidance (**w/o PTA**), randomly mixing the image regions leads to a negligible improvement in accuracy on VQA and NLVR compared to the baseline model mPLUG (**w/o TiMix**). This demonstrates the effectiveness of our TiMix approach in leveraging text guidance for improved performance. In the case denoted as **w/o Mix**, we exclude the mix-based data augmentation method and only retain the PTA task. As presented in Table 4, we observed that utilizing only the PTA task still leads to a notable improvement in performance. This finding suggests that PTA enables our model to learn fine-grained cross-modal semantic alignment, thereby enhancing performance, although the improvement may not be substantial.

Impact of Pre-training Data

To gain a deeper comprehension of the influence of pre-training data size on the efficacy of TiMix , we conducted pre-training with data sizes of 4M, 6M, 8M, 10M, and 14M. These datasets were further augmented using two different mixing strategies: TiMix , Mixup (Zhang et al. 2017) and CutMix (Yun et al. 2019). Figure 4 (a) showcases the VQA results for various mix strategies, as well as the baseline model mPLUG, which does not employ mix-based augmentation.

Models	Pre-train Data	COCO Caption							
		Cross-entropy Optimization				CIDEr Optimization			
		B@4	M	C	S	B@4	M	C	S
E2E-VLP	4M	36.2	-	117.3	-	-	-	-	-
OSCAR	6.5M	36.5	30.3	123.7	23.1	40.5	29.7	137.6	22.8
VinVL	5.65M	38.5	30.4	130.8	23.4	41.0	31.1	140.9	25.2
BLIP	14M	38.6	-	129.7	-	-	-	-	-
LEMON	200M	40.6	30.4	135.7	23.5	42.3	31.2	144.3	25.3
UFO	4M	38.7	30.0	131.2	23.3	-	-	-	-
SimVLM	1.8B	39.0	**32.9**	134.8	24.0	-	-	-	-
mPLUG	4M	39.5	30.9	132.6	23.2	41.4	31.0	140.7	25.3
mPLUG-TiMix	4M	40.7	31.2	134.8	23.9	41.5	30.9	141.2	25.4
mPLUG	14M	41.4	31.0	136.8	23.6	44.8	31.3	148.2	25.8
mPLUG-TiMix	14M	**42.2**	31.4	**138.3**	**24.1**	**45.2**	**31.8**	**151.1**	**26.0**

Table 3: Evaluation Results on image captioning on COCO Karpathy test split (Karpathy and Fei-Fei 2015). B@4: BLEU@4, M: METEOR, C: CIDEr, S: SPICE.

model	PTA	Mix	VQA dev	NLVR dev	PT
mPLUG-TiMix	✓	✓	78.63	84.12	3.5h
-w/o PTA	×	✓	77.89	82.59	3.4h
-w/o Mix	✓	×	78.21	83.13	3.1h
-w/o TiMix	×	×	77.83	82.66	3h

Table 4: The results of the ablation studies where we report the VQA and NLVR performance of various model variants. PT refers to Pretraining Time

Notably, TiMix consistently exhibits superior performance across the entire range. This observation suggests that TiMix not only enhances data efficiency in scenarios with limited data but also delivers substantial performance gains as the dataset size expands. From Figure 4 (a), we have observed that CutMix and Mixup provides only limited improvements in model performance. This indicates that the conventional Mixup approach does not significantly enhance the model's performance. Furthermore, it demonstrates the effectiveness of our approach.

Data Efficiency of TiMix

To explore the effects of TiMix on the additional computational costs during pre-training, we conducted experiments to measure the training time per epoch for mPLUG trained without any data augmentation, as well as mPLUG utilizing TiMix, CutMix (Yun et al. 2019) and Mixup (Yun et al. 2019) for contrastive learning with additional data augmentation. As shown in Figure 4 (b), we recorded the corresponding training times for various data scales. We found that although TiMix , CutMix and mixup introduce some additional training time and computational overhead, the increase in overhead is not significant. Compared to the baseline model mPLUG, mPLUG-TiMix achieves significant improvements in model performance with relatively less computational cost. For example, to achieve the same performance as mPLUG-TiMix on 4M data size, the baseline mPLUG would require scaling the pre-training data to 10M, resulting in much higher computational overhead. This demonstrates the data efficiency of TiMix . Additionally, as shown in Figure 1 (b), we visualize the training loss curves of mPLUG-TiMix trained on 4M data compared to mPLUG trained without TiMix on the same 4M data. We observe that TiMix helps in faster and lower convergence of the training loss. For example, at 30 epochs, the mPLUG-TiMix model has a loss of around 2.1, while the mPLUG model has a loss of around 3.0.

Conclusion

This paper addresses the challenges of scaling up training data volume in Self-supervised Multi-modal Contrastive Learning (SMCL) for Vision-Language Pre-training (VLP) models. We have introduced Text-aware Image Mixing (TiMix) as a solution to improve data efficiency in VLP by integrating mix-based data augmentation techniques into SMCL. Through a theoretical analysis from a mutual information (MI) perspective, we have theoretically shown that well-mixed data samples serve as a regularizer for the classical InfoNCE loss, empirically resulting in significant performance improvements without incurring excessive computational overhead and thereby significantly improving data efficiency in VLP.

Acknowledgments

This research is supported by the National Key Research And Development Program of China (No. 2021YFC3340101).

References

Agrawal, A.; Lu, J.; Antol, S.; Mitchell, M.; Zitnick, C. L.; Parikh, D.; and Batra, D. 2015. VQA: Visual Question Answering. *International Journal of Computer Vision*, 123: 4–31.

Agrawal, H.; Desai, K.; Wang, Y.; Chen, X.; Jain, R.; Johnson, M.; Batra, D.; Parikh, D.; Lee, S.; and Anderson, P. 2018. nocaps: novel object captioning at scale. *CoRR*, abs/1812.08658.

Bi, B.; Li, C.; Wu, C.; Yan, M.; Wang, W.; Huang, S.; Huang, F.; and Si, L. 2020. Palm: Pre-training an autoencod-

ing&autoregressive language model for context-conditioned generation. *arXiv preprint arXiv:2004.07159*.

Changpinyo, S.; Sharma, P.; Ding, N.; and Soricut, R. 2021a. 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*, 3558–3568.

Changpinyo, S.; Sharma, P.; Ding, N.; and Soricut, R. 2021b. 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*, 3558–3568.

Chen, Y.-C.; Li, L.; Yu, L.; Kholy, A. E.; Ahmed, F.; Gan, Z.; Cheng, Y.; and Liu, J. 2020. UNITER: UNiversal Image-TExt Representation Learning. In *ECCV*.

Dabouei, A.; Soleymani, S.; Taherkhani, F.; and Nasrabadi, N. M. 2020. SuperMix: Supervising the Mixing Data Augmentation. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 13789–13798.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *ArXiv*, abs/1810.04805.

Dou, Z.-Y.; Xu, Y.; Gan, Z.; Wang, J.; Wang, S.; Wang, L.; Zhu, C.; Liu, Z.; Zeng, M.; et al. 2021. An Empirical Study of Training End-to-End Vision-and-Language Transformers. *arXiv preprint arXiv:2111.02387*.

Fang, Z.; Wang, J.; Hu, X.; Wang, L.; Yang, Y.; and Liu, Z. 2021. Compressing visual-linguistic model via knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1428–1438.

Goyal, Y.; Khot, T.; Summers-Stay, D.; Batra, D.; and Parikh, D. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6904–6913.

Huang, Z.; Zeng, Z.; Liu, B.; Fu, D.; and Fu, J. 2020. Pixel-BERT: Aligning Image Pixels with Text by Deep Multi-Modal Transformers. *ArXiv*, abs/2004.00849.

Ji, Y.; Tu, R.; Jiang, J.; Kong, W.; Cai, C.; Zhao, W.; Wang, H.; Yang, Y.; and Liu, W. 2023a. Seeing What You Miss: Vision-Language Pre-Training With Semantic Completion Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 6789–6798.

Ji, Y.; Wang, J.; Gong, Y.; Zhang, L.; Zhu, Y.; Wang, H.; Zhang, J.; Sakai, T.; and Yang, Y. 2023b. MAP: Multimodal Uncertainty-Aware Vision-Language Pre-Training Model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 23262–23271.

Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.-T.; Parekh, Z.; Pham, H.; Le, Q. V.; Sung, Y.-H.; Li, Z.; and Duerig, T. 2021. Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. In *International Conference on Machine Learning*.

Jiang, C.; Xu, H.; Ye, W.; Ye, Q.; Li, C.; Yan, M.; Bi, B.; Zhang, S.; Huang, F.; and Huang, S. 2023a. BUS : Efficient and Effective Vision-language Pre-training with Bottom-Up Patch Summarization. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2888–2898.

Jiang, C.; Xu, H.; Ye, W.; Ye, Q.; Li, C.; Yan, M.; Bi, B.; Zhang, S.; Zhang, J.; and Huang, F. 2023b. COPA : Efficient Vision-Language Pre-training through Collaborative Object- and Patch-Text Alignment. *Proceedings of the 31st ACM International Conference on Multimedia*.

Jiang, C.; Ye, W.; Xu, H.; yan, M.; Zhang, S.; Zhang, J.; and Huang, F. 2023c. Vision Language Pre-training by Contrastive Learning with Cross-Modal Similarity Regulation. In *Annual Meeting of the Association for Computational Linguistics*.

Karpathy, A.; and Fei-Fei, L. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3128–3137.

Kim, J.-H.; Choo, W.; Jeong, H.; and Song, H. O. 2021. Co-Mixup: Saliency Guided Joint Mixup with Supermodular Diversity. *ArXiv*, abs/2102.03065.

Kim, W.; Son, B.; and Kim, I. 2021. ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision. In *ICML*.

Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.-J.; Shamma, D. A.; Bernstein, M. S.; and Fei-Fei, L. 2016. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *International Journal of Computer Vision*, 123: 32–73.

Li, C.; Xu, H.; Tian, J.; Wang, W.; Yan, M.; Bi, B.; Ye, J.; Chen, H.; Xu, G.; Cao, Z.; Zhang, J.; Huang, S.; Huang, F.; Zhou, J.; and Si, L. 2022a. mPLUG: Effective and Efficient Vision-Language Learning by Cross-modal Skip-connections.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022b. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv preprint arXiv:2201.12086*.

Li, J.; Selvaraju, R. R.; Gotmare, A. D.; Joty, S. R.; Xiong, C.; and Hoi, S. C. H. 2021. Align before Fuse: Vision and Language Representation Learning with Momentum Distillation. In *NeurIPS*.

Li, L. H.; Yatskar, M.; Yin, D.; Hsieh, C.-J.; and Chang, K.-W. 2019. VisualBERT: A Simple and Performant Baseline for Vision and Language. *ArXiv*, abs/1908.03557.

Li, L. H.; Zhang, P.; Zhang, H.; Yang, J.; Li, C.; Zhong, Y.; Wang, L.; Yuan, L.; Zhang, L.; Hwang, J.-N.; Chang, K.-W.; and Gao, J. 2022c. Grounded Language-Image Pre-training. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Li, X.; Yin, X.; Li, C.; Hu, X.; Zhang, P.; Zhang, L.; Wang, L.; Hu, H.; Dong, L.; Wei, F.; Choi, Y.; and Gao, J. 2020. Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. In *ECCV*.

Liang, W.; Zhang, Y.; Kwon, Y.; Yeung, S.; and Zou, J. Y. 2022. Mind the Gap: Understanding the Modality Gap in

Multi-modal Contrastive Representation Learning. *ArXiv*, abs/2203.02053.

Lin, T.-Y.; Maire, M.; Belongie, S. J.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft COCO: Common Objects in Context. In *ECCV*.

Liu, J.; Liu, B.; Zhou, H.; Li, H.; and Liu, Y. 2022. TokenMix: Rethinking Image Mixing for Data Augmentation in Vision Transformers. In *European Conference on Computer Vision*.

Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; yue Li, C.; Yang, J.; Su, H.; Zhu, J.-J.; and Zhang, L. 2023. Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection. *ArXiv*, abs/2303.05499.

Liu, X.; Zhang, F.; Hou, Z.; Wang, Z.; Mian, L.; Zhang, J.; and Tang, J. 2020. Self-Supervised Learning: Generative or Contrastive. *IEEE Transactions on Knowledge and Data Engineering*, 35: 857–876.

Loshchilov, I.; and Hutter, F. 2019. Decoupled Weight Decay Regularization. In *ICLR*.

Lu, J.; Batra, D.; Parikh, D.; and Lee, S. 2019. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. In *NeurIPS*.

Oord, A. v. d.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Ordonez, V.; Kulkarni, G.; and Berg, T. L. 2011. Im2Text: Describing Images Using 1 Million Captioned Photographs. In *NIPS*.

Plummer, B. A.; Wang, L.; Cervantes, C. M.; Caicedo, J. C.; Hockenmaier, J.; and Lazebnik, S. 2015. Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models. *International Journal of Computer Vision*, 123: 74–93.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021a. Learning Transferable Visual Models From Natural Language Supervision. In *ICML*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021b. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*.

Rennie, S. J.; Marcheret, E.; Mroueh, Y.; Ross, J.; and Goel, V. 2017. Self-Critical Sequence Training for Image Captioning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1179–1195.

Sharma, P.; Ding, N.; Goodman, S.; and Soricut, R. 2018a. Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning. In *Annual Meeting of the Association for Computational Linguistics*.

Sharma, P.; Ding, N.; Goodman, S.; and Soricut, R. 2018b. 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)*, 2556–2565.

Suhr, A.; Zhou, S.; Zhang, A.; Zhang, I.; Bai, H.; and Artzi, Y. 2018. A corpus for reasoning about natural language grounded in photographs. *arXiv preprint arXiv:1811.00491*.

Suhr, A.; Zhou, S.; Zhang, I.; Bai, H.; and Artzi, Y. 2019. A Corpus for Reasoning about Natural Language Grounded in Photographs. *ArXiv*, abs/1811.00491.

Syed, A. A.; Gaol, F. L.; and Matsuo, T. 2021. A Survey of the State-of-the-Art Models in Neural Abstractive Text Summarization. *IEEE Access*, 9: 13248–13265.

Tan, H. H.; and Bansal, M. 2019. LXMERT: Learning Cross-Modality Encoder Representations from Transformers. *ArXiv*, abs/1908.07490.

Uddin, A. F. M. S.; Monira, M. S.; Shin, W.; Chung, T.; and Bae, S.-H. 2020. SaliencyMix: A Saliency Guided Data Augmentation Strategy for Better Regularization. *ArXiv*, abs/2006.01791.

Walawalkar, D.; Shen, Z.; Liu, Z.; and Savvides, M. 2020. Attentive Cutmix: An Enhanced Data Augmentation Approach for Deep Learning Based Image Classification. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 3642–3646.

Wang, W.; Bao, H.; Dong, L.; and Wei, F. 2021a. VLMo: Unified Vision-Language Pre-Training with Mixture-of-Modality-Experts. *ArXiv*, abs/2111.02358.

Wang, Z.; Yu, J.; Yu, A. W.; Dai, Z.; Tsvetkov, Y.; and Cao, Y. 2021b. SimVLM: Simple Visual Language Model Pretraining with Weak Supervision. *ArXiv*, abs/2108.10904.

Xu, H.; Yan, M.; Li, C.; Bi, B.; Huang, S.; Xiao, W.; and Huang, F. 2021. E2E-VLP: End-to-End Vision-Language Pre-training Enhanced by Visual Learning. *ArXiv*, abs/2106.01804.

Yao, L.; Huang, R.; Hou, L.; Lu, G.; Niu, M.; Xu, H.; Liang, X.; Li, Z.; Jiang, X.; and Xu, C. 2021. FILIP: Fine-grained Interactive Language-Image Pre-Training. *arXiv preprint arXiv:2111.07783*.

Yu, F.; Tang, J.; Yin, W.; Sun, Y.; Tian, H.; Wu, H.; and Wang, H. 2021. ERNIE-ViL: Knowledge Enhanced Vision-Language Representations Through Scene Graph. In *AAAI*.

Yu, L.; Poirson, P.; Yang, S.; Berg, A. C.; and Berg, T. L. 2016. Modeling context in referring expressions. In *European Conference on Computer Vision*, 69–85. Springer.

Yun, S.; Han, D.; Oh, S. J.; Chun, S.; Choe, J.; and Yoo, Y. J. 2019. CutMix: Regularization Strategy to Train Strong Classifiers With Localizable Features. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 6022–6031.

Zeng, Y.; Zhang, X.; and Li, H. 2021. Multi-Grained Vision Language Pre-Training: Aligning Texts with Visual Concepts. *ArXiv*, abs/2111.08276.

Zhang, H.; Cissé, M.; Dauphin, Y.; and Lopez-Paz, D. 2017. mixup: Beyond Empirical Risk Minimization. *ArXiv*, abs/1710.09412.

Zhang, P.; Li, X.; Hu, X.; Yang, J.; Zhang, L.; Wang, L.; Choi, Y.; and Gao, J. 2021. VinVL: Revisiting Visual Representations in Vision-Language Models. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5575–5584.