RILS: Masked Visual Reconstruction in Language Semantic Space

Shusheng Yang[1,*], Yixiao Ge[2,†], Kun Yi[2], Dian Li[3], Ying Shan[2], Xiaohu Qie[4], Xinggang Wang[1,†]
[1]School of EIC, Huazhong University of Science & Technology
[2]ARC Lab, [4]Tencent PCG [3]Foundation Technology Center, [4]Tencent PCG
{shushengyang, xgwang}@hust.edu.cn {yixiaoge, kunyi, goodli, yingsshan, tigerqie}@tencent.com

Abstract

Both masked image modeling (MIM) and natural language supervision have facilitated the progress of transferable visual pre-training. In this work, we seek the synergy between two paradigms and study the emerging properties when MIM meets natural language supervision. To this end, we present a novel masked visual Reconstruction In Language semantic Space (RILS) pre-training framework, in which sentence representations, encoded by the text encoder, serve as prototypes to transform the vision-only signals into patch-sentence probabilities as semantically meaningful MIM reconstruction targets. The vision models can therefore capture useful components with structured information by predicting proper semantic of masked tokens. Better visual representations could, in turn, improve the text encoder via the image-text alignment objective, which is essential for the effective MIM target transformation. Extensive experimental results demonstrate that our method not only enjoys the best of previous MIM and CLIP but also achieves further improvements on various tasks due to their mutual benefits. RILS exhibits advanced transferability on downstream classification, detection, and segmentation, especially for low-shot regimes. Code is available at https://github.com/hustvl/RILS.

1. Introduction

Learning transferable representation lies a crucial task in deep learning. Over the past few years, natural language processing (NLP) has achieved great success in this line of research [18, 45, 60]. To explore similar trajectories in the vision domain, researchers tend to draw upon the successes of NLP and have made tremendous progress: (i) Inspired by the advanced model architecture [60] as well as self-supervised learning paradigm [18] in NLP, vision Transformers (ViT) [21, 42] and masked image modeling (MIM) [3, 28] open a new era of self-supervised visual representation learning, and show unprecedented transferability on various tasks, especially on fine-grained tasks such as object detection [22, 39]. (ii) Inspired by the great scalability brought by leveraging web-scale collections of texts as training data in NLP [5, 48, 49, 72], CLIP [47] and ALIGN [35] bring such a principle to vision and manifest the immense potential of leveraging natural language supervision for scalable visual pre-training. Strong transferability of pre-trained visual models on low-shot regimes ensues. It also facilitates diverse applications by extracting contextualized image or text features [26, 50, 52].

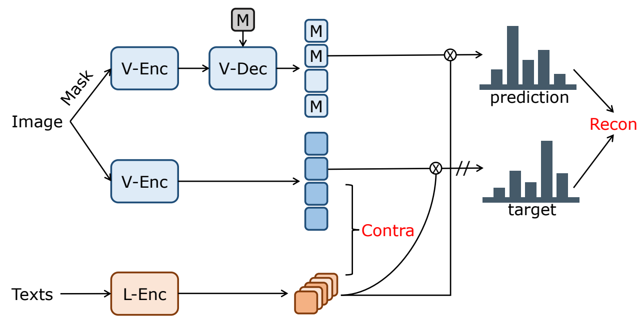

Figure 1. **Overview of our RILS.** Recon and Contra represent masked reconstruction loss and image-text contrastive loss. During pre-training, RILS learns to perform masked image modeling and image-text contrastive simultaneously. Masked predictions and corresponding targets are transformed to probabilistic distributions in language space by leveraging text features as semantic-rich prototypes. Under such a scheme, both objective are unified and achieve mutual benefits from each other. Vision encoder obtains the ability to capture meaningful and fine-grained context information, sparking decent transfer capacity.

The remarkable attainment achieved by these two research lines pushes us to ponder: *Is it possible to unify both masked image modeling and natural language supervision to pursuit better visual pre-training?* A straightforward way towards this goal is to simply combine masked image modeling (MIM) with image-text contrastive learning (ITC) for multi-task learning. Although the naïve combination is feasible to inherit the above advantages, we find it remains un-

[*]This work was done while S. Yang was an intern at ARC Lab, Tencent PCG. [†] X. Wang and Y. Ge are the corresponding authors.

satisfactory due to the mutual benefits between MIM and ITC have not yet been fully explored. Motivated by this, we develop RILS, a tailored framework to seek the synergy of masked image modeling and language supervision.

The core insight of RILS is to perform *masked visual reconstruction in language semantic space*. Specifically, instead of reconstructing masked patches in the standalone vision space (*e.g.*, raw pixel [28,66], low-level features [3,62] or high-level perceptions [12, 34, 63, 75]), we map patch features to a probabilistic distribution over a batch of text features as the reconstruction target, which is enabled by ITC that progressively aligns the image and text spaces. The text features serve as semantically rich prototypes and probabilistic distributions explicitly inject the semantic information onto each image patch. The MIM objective is formulated as a soft cross-entropy loss to minimize the KL divergence between text-injected probabilistic distributions of masked vision tokens and their corresponding targets. The visual model optimized by our language-assisted reconstruction objective, in turn, improves ITC with better visual representations that capture fine-grained local contexts.

Under such a working mechanism, the two objectives(*i.e.*, MIM and ITC) complement each other and form a unified solution for transferable and scalable visual pre-training. Note that a lot of works [34, 63, 75] have manifested the importance of semantic information in the reconstruction target of MIM objectives. However, it is abstract to pursue such a semantically rich space with visual-only signals due to its unstructured characteristics [29]. Thanks to natural language supervision, this issue is alleviated by performing masked reconstruction in language space.

Extensive experiments on various downstream tasks demonstrate that our design enjoys the best of both worlds. With a vanilla ViT-B/16 as the vision model and 25-epoch pre-training on 20 million image-text pairs, RILS achieves 83.3% top-1 accuracy when fine-tune on ImageNet-1K [15], $+1.2\%$ and $+0.6\%$ better than the MAE [28] and CLIP [47] counterparts. Advanced performance can be consistently acquired when transfer to fine-grained tasks such as detection and segmentation. Moreover, our approach exhibits promising out-of-the-box capability under an extremely low-shot regime. RILS also demonstrates superior performance on zero-shot image classification and image-text retrieval. On ImageNet-1K benchmark, RILS obtains 45.0% zero-shot classification accuracy, $+4.7\%/+3.4\%$ higher than CLIP [47]/SLIP [43] under the same training recipe. Compelling results of RILS imply the promising capacity in the unification of MIM and language supervision.

2. Related Works

Masked Image Modeling translates masked language modeling [18] to vision domain and learns transferable visual representation by reconstructing masked signals from partial observation [3, 9, 21]. Despite following the same *mask-then-reconstruction* principle, MIM differs from MLM a lot in the design of reconstruction target. BEiT [3] utilizes a pre-trained d-VAE [51, 55] and reconstructs masked image patches in the offline token space. Subsequent works improve it by employing better pre-trained tokenizer [19, 34, 44, 63], eased and refined multi-choice tokens [37] or contextual aligner [11]. MAE [28] and SimMIM [66] demonstrate directly reconstruct masked patches in raw pixel space can also lead to favorable transferability as well as scalability. MaskFeat [62] takes hand-crafted low-level HOG feature [14] as target. Other works like iBOT [75], data2vec [2] and SdAE [12] perform reconstruction in a high-level vision feature space. Different from these methods, in this work, we tap the potential when masked image modeling meets natural language supervision and propose performing masked visual reconstruction in the language semantic space.

Language Supervised Visual Pre-training learns visual representation from image-text pairs by solving generative [16,56] or discriminative [73] pretext tasks. Recently, benefit from modern network architectures [21,41,42] and publicly available image-text datasets [8,17,57–59], CLIP [47] and ALIGN [35] unveil the tremendous transferability and scalability of this paradigm. The core technique of CLIP is aligning both vision and language modalities in a joint embedding space by global representation contrastive. Follow-up works further improve CLIP on the vision-only [43, 67] or vision-language [38, 64, 68–70] side. In this paper, we bring natural language supervision together with masked image modeling for better visual pre-training on these two paradigms.

3. Our Approach

3.1. Architecture

Among numerous architecture designing spaces, without loss of generalization, we adopt an asymmetric *encoder-decoder* architecture following MAE [28] and a *dual-encoder* architecture following CLIP [47] for their flexibility. As illustrated in Figure 1, RILS comprises the following three major components:

Vision Encoder plays the key role in RILS and all our efforts aim to strengthen its capacity on downstream transfer. Following the trend in recent visual pre-training, we implement this encoder by a vanilla vision Transformer (ViT) [21]. It takes both intact (unmasked) image and corrupted (masked) image as inputs. Formally, input image I is first divided into regular non-overlapping image patches and then encoded by a stack of Transformer blocks [60]. Meanwhile, following MAE [28], we randomly mask a large portion of image patches and leave the remaining patches to be visible. This corrupted image $\hat{I}$ is also encoded by vision encoder. We formulate the process of vision encoder as:

$$\begin{aligned}&\text{V-Enc}(I) = \{f^k | k \in [1, N]\},\\ &\text{V-Enc}(\hat{I}) = \{\hat{f}^k | k \in [1, N] \backslash \mathcal{M}\},\end{aligned} \tag{1}$$

in which k denotes the patch index and N denotes image patch numbers. f and $\hat{f}$ betoken encoded patch features of intact image I and corrupted image $\hat{I}$. $\mathcal{M}$ indicates the index set of random masked patches.

Language Encoder encodes input text T by a stack of Transformer layers with causal masked attention [60]. This process can be simply represented by:

$$\text{L-Enc}(T) = h. \tag{2}$$

We take the output h as global representation of input text.

Vision Decoder consists of another series of Transformer blocks. Particularly, in our design, decoder blocks have the same dimension as encoder blocks. It takes the encoded feature $\hat{f}$ of masked image $\hat{I}$ along with a learnable $[\text{MASK}]$ token as inputs, and tries to predict masked signals from corrupted view:

$$\text{V-Dec}(\{\hat{f}^k | k \in [1, N] \backslash \mathcal{M}\}, [\text{MASK}]) = \{g^k | k \in [1, N]\}. \tag{3}$$

3.2. Training Objective

Image-Text Contrastive. We leverage image-text contrastive loss to align two modalities into a joint embedding space. Specifically, given image-text pair $\{(I, T)\}$, we take the mean-pooled image feature $\bar{f} = \frac{1}{N}\sum_{k=1}^{N} f^k$ and h in Eq. (2) as global representations for image and text. The image and text features are further projected by two projection heads and followed by a normalization:

$$\begin{aligned}z^I &= ||\theta(\bar{f})||,\\ z^T &= ||\phi(h)||,\end{aligned} \tag{4}$$

$\theta(\cdot)$ and $\phi(\cdot)$ denotes the projection head for image and text respectively. The image-to-text contrastive loss and text-to-image contrastive loss can be represented as:

$$\begin{aligned}\mathcal{L}_{\text{I2T}} &= -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp(\langle z_i^I, z_i^T\rangle/\sigma)}{\sum_{j=1}^{B}\exp(\langle z_i^I, z_j^T\rangle/\sigma)},\\ \mathcal{L}_{\text{T2I}} &= -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp(\langle z_i^T, z_i^I\rangle/\sigma)}{\sum_{j=1}^{B}\exp(\langle z_i^T, z_j^I\rangle/\sigma)},\end{aligned} \tag{5}$$

in which i and j stands for the index within a mini-batch and B indicates the batch size respectively. σ performs a learnable temperature and is jointly trained during the pre-training. The total loss of image-text contrastive learning can be formulated as:

$$\mathcal{L}_{\text{Contra}} = \frac{1}{2}(\mathcal{L}_{\text{I2T}} + \mathcal{L}_{\text{T2I}}). \tag{6}$$

Masked Visual Reconstruction in Language Semantic Space. As aforementioned, despite the *mask-then-reconstruct* principle of MIM is concise enough, the contiguous and unstructured characteristics in visual signal make the choice of reconstruction space non-trivial. Lots of works have manifested the great importance of performing masked reconstruction in a semantic-rich space [12, 34, 63, 75]. In this work, we build our reconstruction space from a vision-language perspective. We regard the text features as natural semantic descriptors for image patches and try to perform masked visual reconstruction in this language space. Specifically, given the encoded patch feature f^k in Eq. (1) with k being the index of patch and decoded patch feature g^k in Eq. (3), we firstly project and normalize both features to the vision-language aligned space:

$$\begin{aligned}\tilde{f}_i^k &= ||\theta(f_i^k)||,\\ \tilde{g}_i^k &= ||\theta(g_i^k)||,\end{aligned} \tag{7}$$

with i being the index within mini-batch. $\theta(\cdot)$ represents the same vision projection head in Eq. (4). The key step of our design is to map patch features to a probabilistic distributions over a bunch of text features:

$$\begin{aligned}\boldsymbol{p}_i^k &= \{\frac{\exp(\langle \tilde{f}_i^k, z_l^T\rangle/\tau_1)}{\sum_{j=1}^{B}\exp(\langle \tilde{f}_i^k, z_j^T\rangle/\tau_1)} \mid l \in [1, B]\},\\ \boldsymbol{q}_i^k &= \{\frac{\exp(\langle \tilde{g}_i^k, z_l^T\rangle/\tau_2)}{\sum_{j=1}^{B}\exp(\langle \tilde{g}_i^k, z_j^T\rangle/\tau_2)} \mid l \in [1, B]\},\end{aligned} \tag{8}$$

in which τ_1 and τ_2 serve as temperatures. In this way, with the text features serve as semantic-rich prototypes, both masked prediction and corresponding target are mapped into this language semantic space. The probabilistic distributions explicitly express the context information within each patch. The reconstruction objective is to shrinking the differences between text-injected distributions of target and masked prediction by minimize the KL divergence of p_i^k w.r.t. q_i^k, which can be represented by:

$$\mathcal{L}_{\text{Recon}} = \frac{1}{||\mathcal{C}|| \cdot ||\mathcal{M}||}\sum_{i\in\mathcal{C}}\sum_{k\in\mathcal{M}} -\text{sg}[\boldsymbol{p}_i^k]\log \boldsymbol{q}_i^k, \tag{9}$$

in which $\text{sg}[\cdot]$ indicates stop gradient. $\mathcal{M}$ denotes the index set of masked patches. $\mathcal{C}$ signifies the indexes of images which are correctly aligned to corresponding text features. In other words, we only calculate reconstruction loss on images which are correctly matched with target texts in image-to-text matching.

By transferring reconstruction space from standalone vision space to language space, our approach takes both MIM and ITC into a unifying landscape and achieves mutual benefits from each other. MIM always suffers from overly paying attention on low-level details which consume lots of

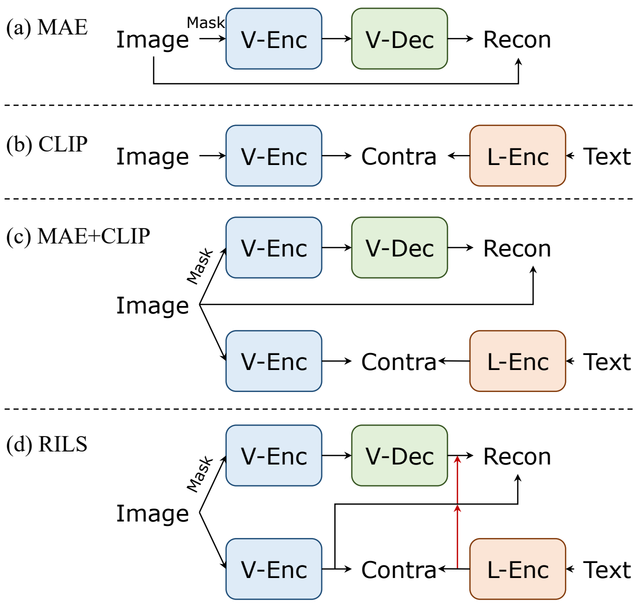

Figure 2. Architecture comparisons between MAE [28], CLIP [47], MAE+CLIP and RILS. Recon and Contra indicate masked reconstruction loss and image-text contrastive loss.

model's capacity but of less helpful for understanding visual concepts. By leveraging text features as prototypes and transfer patch features to probabilistic distribution on language space, the low-level information inside visual signals are abandoned by the contextualized language prototypes. Better contextualized image features in turn assist vision-language alignment, leading to transferable and scalable visual pre-training.

Overall Objective Function. The final objective of RILS is a weighted-sum of both image-text contrastive loss and masked reconstruction loss:

$$\mathcal{L}_{\text{RILS}} = \lambda_1 \cdot \mathcal{L}_{\text{Contra}} + \lambda_2 \cdot \mathcal{L}_{\text{Recon}}. \tag{10}$$

λ_1 and λ_2 indicate coefficients to balance two losses.

3.3. Pre-training Setup

Similar to [69], we sequentially sample image-text pairs according to filenames from recent released LAION-400M [58] dataset as our training sets. We term them as L-10M/L-20M/L-50M according to the amount of sampled unique image-text pairs (*e.g.*, L-10M stands for the first 10 million subset of LAION-400M). Unless specified, our method is trained on L-20M for 25 epochs. We take AdamW [36] as optimizer with learning rate set to 5e-4 and a weight decay of 0.5. Learning rate linearly increases in the first epoch as warmup and decreases in the rest following the cosine learning rate decay strategy. We train our method on 32 NVIDIA V100 with a total batch size of 4096 (*i.e.*, batch size per GPU is 128). For model architecture, we take the widely-adopted ViT-B/16 as vision encoder, 1-layer Transformer block with 768-dim and 12 heads as vision decoder, and a text Transformer with 12 blocks and 512-dim as language encoder. To tokenize text inputs, following [47], we use byte-pair encoding (BPE [23]) with 49K vocabulary size and set the max length of each sentence to 77. During pre-training, input images are resized to 224×224 and we set random mask ratio to 75% following [28]. Temperatures τ_1 and τ_2 in Eq. (8) are set to 0.04 and 0.1. Loss coefficients λ_1 and λ_2 are set to 1.0 and 0.5 by default. More pre-training setups are listed in the appendix.

Method	Dataset	PT Epo.	Lin.	FT.
SimCLR [10]	L-20M	25(∼ 400)	51.7	81.3
MAE [28]			44.3	82.1
CLIP [47]			67.8	82.7
SLIP [43]			70.1	82.6
MAE+CLIP			64.5	82.9
RILS			**71.5**	**83.3**
BEiT [3]	IN-1K(∼ 1.3M)	800	–	83.2
MAE [28]		1600	67.8	83.6
RILS	L-50M	25(∼ 1000)	**71.9**	**83.6**

Table 1. **Image classification results on ImageNet-1K (IN-1K).** PT Epo. indicates per-training epochs. Lin. and FT. is short for linear probing and end-to-end fine-tuning respectively.

3.4. Discussion

There are also meaningful attempts [20,34,44,63] on utilizing natural language supervision together with MIM and seem alike to ours. However, there still have some distinctions existing in the motivation and method between ours and theirs. MVP [63], MILAN [34] and BEiTv2 [44] leverage natural language supervision by a two-stage framework, while ours is fully end-to-end. We will take further discussion about two-stage methods and ours in later experiments. A related and concurrent work [20] shows some similar designs but significantly differs from our idea of reconstruction in language space. As it does not have a reproducible implementation, we do not take it into consideration for comparison.

4. Main Results

In this section, we evaluate the representation quality of pre-training by transferring pre-trained models to various downstream tasks. We choose MAE [28] and CLIP [47] as representative methods of masked image modeling and vision-language contrastive learning. We also conduct a naïve baseline (termed as MAE+CLIP) which simply combine MAE and CLIP together to perform MIM and ITC simouteneously, as a counterpart for our approach. Semantically comparisons between MAE, CLIP, MAE+CLIP and our RILS are illustrated in Figure 2.

Method	COCO						LVIS					
	$AP^{\mathcal{B}}$	$AP^{\mathcal{B}}_{50}$	$AP^{\mathcal{B}}_{75}$	$AP^{\mathcal{M}}$	$AP^{\mathcal{M}}_{50}$	$AP^{\mathcal{M}}_{75}$	$AP^{\mathcal{B}}$	$AP^{\mathcal{B}}_{50}$	$AP^{\mathcal{B}}_{75}$	$AP^{\mathcal{M}}$	$AP^{\mathcal{M}}_{50}$	$AP^{\mathcal{M}}_{75}$
MAE [28]	48.1	68.6	52.9	42.4	65.8	**46.4**	31.0	46.2	33.7	29.6	44.0	31.7
CLIP [47]	47.7	69.1	52.3	42.0	65.9	45.1	32.3	48.1	35.1	30.5	45.9	32.5
SLIP [43]	46.5	68.5	51.0	41.5	65.1	44.1	32.4	48.9	35.1	30.8	46.5	32.6
MAE+CLIP	48.1	69.6	52.5	42.4	66.2	45.7	32.6	48.8	35.2	30.7	46.4	32.6
RILS	**48.5**	**70.5**	**53.2**	**42.6**	**66.8**	45.8	**33.8**	**50.2**	**36.5**	**31.6**	**47.8**	**33.7**

Table 2. **Object detection and instance segmentation results on COCO and LVIS.** All models are pre-trained with ViT-B/16 for 25 epochs on L-20M. Fine-tuning recipes for different pre-trained models are the same.

4.1. Classification Transfer

Linear Probing evaluates the quality of pre-trained feature by training a linear classifier on the frozen feature. We following the the recipe in [7] and sweep over different learning rate. Results are shown in Table 1. We notice that, with 25-epochs pre-training on L-20M, MAE+CLIP only achieves 64.5% accuracy which is better than MAE but worse than CLIP. This implies the contradiction between MIM and ITC exists in such a naïve combination. RILS alleviates this contradiction by elaborated design and outperforms other methods by a large margin.

End-to-End Fine-tuning. We follow most of setups in [28] for fine-tuning. Concretely, we fine-tune pre-trained models for 100 epochs with a warm-up of 5 epochs. Hyper-parameters are all the same for all experiments except the learning rate. Results are shown in Table 1. When training with 25 epochs on L-20M, our method shows distinct advantages. Compared to MAE, CLIP and MAE+CLIP, our method exhibits +1.2%, +0.6% and +0.4% gains respectively. Moreover, when we scale-up the dataset capacity from L-20M to L-50M, with 25 epochs pre-training (around 1000 equivalent epochs in the ImageNet-1K regime), our method achieves 83.6% top-1 accuracy which is on par with prior art (MAE trained on ImageNet-1K with 1600 epochs) with only 62.5% training length.

4.2. Downstream Transfer

Object Detection and Segmentation. For object detection and instance segmentation, we choose COCO [40] and LVIS [27] as benchmarks. We follow the design in [39] to transfer pre-trained ViT to detection. To tame quadratic complexity within self-attention, most attention blocks in the ViT are replaced with window attention except for four global blocks to perform cross-window interaction. SimpleFPN [39] is attached to the last transform block to generate pyramid features. Modernized RPN [54] and Mask R-CNN [31] head are deployed for detecting and segmenting visual instances. All pre-trained models are fine-tuned on two benchmarks for 25 epochs with same hyper-parameters. The results are shown in Table 2. Among all methods, our RILS achieves the best results in terms of $AP^{\mathcal{B}}$ and $AP^{\mathcal{M}}$ on both COCO and LVIS. It's noteworthy that two benchmarks show different properties: COCO benchmark shows less benefits from language supervision while LVIS converses. Specifically, MAE shows leading performance on COCO but inferior performance on LVIS. We suspect this is due to the inherent distinctions in COCO and LVIS: LVIS contains 1203 visual categories which is about 15× more than COCO, and it always suffers from the long-tail distribution. Under such circumstances, COCO requires more localization ability which MAE excel at while LVIS prefers better classification ability which natural language supervision can bring. From this perspective, when compare our RILS with the MAE+CLIP, we find our design benefits more from both MIM and ITC objectives. On both COCO and LVIS, MAE+CLIP only shows competitive performance to the winner of MAE and CLIP, but our RILS exhibits apparent improvements especially on LVIS. This indicates our design leverage masked image modeling and language supervision in a more synergistic way. We believe this kind of synergy is of great exploration value for better visual pre-training.

Method	Dataset	PT Epo.	mIoU
BEiT [3]	IN-1K(∼ 1.3M)	300	45.5
		800	46.5
MAE [28]		300	45.8
		1600	48.1
MAE [28]	L-20M	25(∼ 400)	44.2
CLIP [47]			45.2
SLIP [43]			45.7
MAE+CLIP			45.3
RILS	L-20M	25(∼ 400)	**48.1**

Table 3. **Semantic segmentation results on ADE20K.**

Semantic Segmentation. Experiments on semantic segmentation are conducted on the well-known ADE20K [74] dataset. We build the segmentation framework upon UperNet [65] and use the pre-trained models as encoders. Input images are resized to 512 × 512 and all models are fine-tuned for 160K iterations. All hyper-parameters strictly follow MAE [28] and not tuned. We report the mean intersection-over-union (mIoU) in Table 3. As the results shown, our method overwhelmingly surpasses others. Specifically, with 25 epoch pre-training on L-20M, our method achieves 48.1 mIoU, +3.9 and +2.9 higher than MAE and CLIP. Similar to the trends on LVIS, MAE+CLIP

Method	PT Dataset	PT Epo.	Images per Class			
			1	2	5	10
MAE [28]	IN-1K($\sim$ 1.3M)	1600	4.3 $\pm$ (0.28)	10.6 $\pm$ (0.21)	22.4 $\pm$ (0.28)	31.6 $\pm$ (0.02)
BEiT [3]	IN-1K($\sim$ 1.3M)	800	1.3 $\pm$ (0.03)	2.2 $\pm$ (0.13)	4.4 $\pm$ (0.21)	7.4 $\pm$ (0.05)
MAE [28]	L-20M	25($\sim$ 400)	3.4 $\pm$ (0.12)	5.2 $\pm$ (0.21)	10.1 $\pm$ (0.10)	14.8 $\pm$ (0.20)
CLIP [47]	L-20M	25($\sim$ 400)	19.4 $\pm$ (0.18)	29.2 $\pm$ (0.61)	39.8 $\pm$ (0.39)	46.3 $\pm$ (0.15)
SLIP [43]	L-20M	25($\sim$ 400)	17.7 $\pm$ (0.33)	27.2 $\pm$ (0.56)	38.6 $\pm$ (0.55)	46.4 $\pm$ (0.06)
MAE+CLIP	L-20M	25($\sim$ 400)	21.1 $\pm$ (0.12)	31.1 $\pm$ (0.94)	41.6 $\pm$ (0.43)	47.5 $\pm$ (0.15)
RILS	L-20M	25($\sim$ 400)	**24.0** $\pm$ (0.27)	**34.6** $\pm$ (0.88)	**45.7** $\pm$ (0.46)	**51.8** $\pm$ (0.18)

Table 4. **Extreme low-shot classification on ImageNet-1K.** We random sample 1, 2, 5, 10 images per class from training split, and report logistic regression accuracy (%) on ImageNet-1K validation split. All methods use ViT-B/16 as vision encoder.

only gets 0.1 mIoU gains by simply combining MIM with ITC together, far less than our approach. Furthermore, our method shows competitive or better performance when compared to prior art. Compared to MAE pre-trained on ImageNet-1K with 300 epochs, our method achieves 2.3 higher performance (48.1 *vs.* 45.8). When MAE is pre-trained with 1600 epochs, our method achieves the same mIoU (48.1) while only requires 25% training length.

Experiments above demonstrate the excellent transfer capacity of our approach on fine-grained visual understanding tasks. Our design unleashes the ability to capture local details and global contexts by performing masked visual reconstruction in language semantic space.

4.3. Label-Efficient Transfer

Low-shot Regime Classification. We investigate the classification performance when only very few labeled images are available. Specifically, following [1,6], we random sample 1, 2, 5, and 10 labeled images per class from ImageNet-1K training split as our training sets. Instead of end-to-end fine-tuning, we train a linear classifier on frozen features to avoid overfitting. The complete validation split of ImageNet-1K which contains 50K images are used to evaluate the accuracy. Table 4 shows the results.

Specifically, compared to MAE+CLIP which only obtain slightly improvements over CLIP, our RILS outperforms both of them by a large margin. Notably, with only 10 images per class (*i.e.*, 10K images for 1K classes), our method can achieve 51.8% top-1 accuracy.

Low-shot Regime Detection. We further transfer the low-shot experiment to object detection on COCO [40], which requires model to localize and classify visual objects simultaneously. We randomly sample annotated images from COCO training split with different sampling ratio (range from 1% to 50%) as our training sets. All models are end-to-end fine-tuned for 12 epochs instead of 25 to prevent overfitting. We report the average precision of detection $\mathrm{AP}^{\mathcal{B}}$ for comparison in Table 5. As the results shown, our method shows the best performance under a wide range of sampling ratio (from 2% to 50%).

Method	COCO Sampling Ratio					
	1%	2%	5%	10%	20%	50%
MAE [28]	0.94	6.10	15.76	23.16	29.78	38.10
CLIP [47]	0.81	5.05	14.98	22.49	29.88	38.50
SLIP [43]	**1.11**	4.54	13.84	21.91	29.53	37.73
MAE+CLIP	0.68	5.28	14.33	23.72	29.99	39.24
RILS	0.86	**6.46**	**16.94**	**24.69**	**31.97**	**40.41**

Table 5. **Low-shot regime object detection on COCO**. We report detection performance $\mathrm{AP}^{\mathcal{B}}$ with 12 epochs fine-tuning. All models are pre-trained with ViT-B/16 and 25 epochs on L-20M.

Conceptually, one of the aspirations of pre-training is to pursue efficient transfer (*e.g.*, less trainable data, shorter training length) on downstream tasks [1, 25, 30]. Experiments in the low-shot regime show the strong out-of-the-box capacity of our RILS by performing MIM in language semantic space. The non-trivial results indicate our pre-training approach brings out label-efficient learner, showing great application value to real-world scenarios, especially when annotated data is insufficient.

4.4. Zero-Shot Transfer

Classification. We evaluate the zero-shot classification over 21 benchmarks including ImageNet-1K [15]. Detail of each datasets are listed in the appendix and the evaluate recipes (*e.g.*, prompt engineering) strictly follow [43]. Results are shown in Table 6. Specifically, compared to CLIP, MAE+CLIP only achieves +0.6% average improvements, while our RILS shows +2.0% gains. This hints the masked image modeling objective in the naïve combination has little help to image-text alignment, while ours alleviate this issue by bind two objectives in a unified landscape. On ImageNet-1K, our method achieves 45.0% accuracy, $+4.7\%/+3.4\%/+2.7\%$ higher than CLIP, SLIP and MAE+CLIP, respectively. Among 21 benchmarks, our method outperforms others over 17 datasets, frequently with a significant margin.

Image-Text Retrieval. We study image-text retrieval on 2 benchmarks: COCO [40] and Flickr30K [46]. For both benchmarks, we use the original captions (w/o prompt) and 224×224 resized images for retrieval. Different from zero-

Method	Food101	CIFAR10	CIFAR100	CUB200	SUN397	Cars	Aircraft	DTD	Pets	Caltech101	Flowers	MNIST	FER2013	STL10	EuroSAT	RESISC45	GTSRB	Country211	CLEVR	SST2	ImageNet	**Average**	**# Wins.**
CLIP [47]	55.7	76.0	46.9	**24.4**	50.7	17.8	4.8	31.5	53.7	78.4	31.8	26.8	37.6	89.0	22.7	36.9	**24.1**	6.8	20.0	49.1	40.3	39.3	2
SLIP [43]	56.7	73.4	43.2	22.6	51.6	17.7	4.9	32.4	52.5	79.1	33.3	**29.4**	33.5	89.5	17.8	36.2	17.8	6.8	**23.4**	49.7	41.6	38.7	2
MAE+CLIP	57.8	78.2	52.4	23.9	51.6	18.1	4.6	31.5	55.8	78.4	32.0	27.6	32.7	89.8	27.0	39.4	22.9	7.2	14.7	49.3	42.3	39.9	0
RILS	**58.9**	**86.2**	**55.1**	23.4	**51.8**	**19.5**	**5.9**	**32.8**	**59.2**	**80.7**	**33.5**	22.6	**40.1**	**93.2**	**28.8**	**40.2**	19.1	**7.8**	16.8	**50.0**	**45.0**	**42.3**	**17**

Table 6. **Zero-shot classification on 21 datasets.** All models are trained with ViT-B/16 encoder for 25 epochs on L-20M.

Method	COCO						Flickr30K					
	I→T			T→I			I→T			T→I		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
CLIP [47]	41.82	69.50	79.34	30.54	57.10	69.30	28.13	50.77	61.12	20.61	40.77	50.49
SLIP [43]	44.54	72.20	82.10	33.26	59.66	71.14	31.20	55.81	66.02	24.05	45.58	55.90
MAE+CLIP	42.72	70.66	80.36	31.40	57.50	69.68	28.62	52.25	62.64	22.64	43.67	53.97
RILS	45.06	73.38	83.36	34.86	61.36	72.78	32.21	56.39	66.67	25.48	47.55	57.94

Table 7. **Zero-shot image-text retrieval.** I→T and T→I indicate image-to-text retrieval and text-to-image retrieval.

Method	IN-A	IN-R	IN-Ske	IN-V2	ObjNet	**Average**
CLIP [47]	9.3	51.2	28.1	39.8	17.7	32.3
SLIP [43]	10.5	49.8	26.7	41.3	20.4	$33.1_{\uparrow 0.8}$
MAE+CLIP	11.6	53.9	31.1	41.6	19.4	$34.4_{\uparrow 2.1}$
RILS	**12.1**	**55.7**	**31.4**	**43.3**	**21.0**	$\mathbf{35.7}_{\uparrow 3.4}$

Table 8. **Zero-shot out-of-distribution classification.** We report the Top-1 classification accuracy (%) for reference.

shot classification, retrieval requires the models to have the ability to capture fine-grained contextualized information in both images and texts. As the results in Table 7 shown, our model achieves the best among all counterparts. In particular, when compared to CLIP, our RILS shows more significant improvements than MAE+CLIP.

Robustness to Distribution Shift. Following CLIP [47], we also conduct zero-shot experiments on 5 out-of-distribution datasets: ImageNet Adversarial [33], ImageNet Rendition [32], ImageNetV2 [53], ImageNet Sketch [61], and ObjectNet [4]. As shown in Table 8, our method outperforms others on all five benchmarks. Specifically, RILS obtains significant improvements, +3.4%, +2.6%, +1.3% better than CLIP, SLIP, and MAE+CLIP respectively.

5. Ablation Study

In this section, we ablate the designs of RILS. All experiments are conducted with ViT/B-16 vision encoder and trained on L-10M for 25 epochs. We report the classification accuracy (%) under zero-shot (ZS.), linear probing (Lin.) and end-to-end fine-tuning (FT.) on ImageNet-1K.

5.1. Comparisons with Two-stage Methods

As discussed above, another way to leverage both masked image modeling with image-text contrastive learning is to build a two-stage framework and learn two objectives step by step. In this section, we compare our RILS with several two-stage methods:

MIM→LiT indicates firstly pre-train with masked image modeling only, then follow [71] to perform locked-image text tuning on image-text pairs. Specifically, we start the second stage by fine-tuning pre-trained MAE [28].

MIM→CLIP denotes fully fine-tune pre-trained MAE on image-text pairs in the second stage. In this way, pre-trained model also inherit properties from both objectives.

CLIP→MIM stands for using pre-trained CLIP as a guidance for masked image modeling in the second stage. This paradigm has been studied in recent research such as [34, 44, 63].

The comparison results are shown in Table 9. As the vision encoder is fully frozen during the second stage in MIM→LiT, its performance on downstream tasks remains unchanged except for zero-shot. MIM→CLIP slightly outperforms MAE and CLIP. CLIP→MIM exhibits more improvements upon two base methods, but lose the ability on zero-shot classification. Our method rivals all counterparts with a more concise training pipeline.

5.2. Comparisons on Reconstruction Space

The core philosophy of our design is to perform masked reconstruction in language semantic space. We ablate the effectiveness of our design by comparing to two other alternatives: raw pixel space and high-level vision space. Reconstruction in raw pixel space denotes the aforementioned MAE+CLIP which tries to reconstruct raw pixels directly. For high-level vision space, we replace the language feature z^T in Eq. (8) to learnable weights with other components unchanged. In other words, similar to design in [1,7,75], we map patch features to a probabilistic distribution on a group of learnable weights. As results in Table 10 shown, our

Method	ZS.	Lin.	FT.
MAE [28]	–	43.4	81.5
CLIP [47]	32.1	64.1	82.0
MIM→LiT [71]	13.2	43.4	81.5
MIM→CLIP	34.4	64.8	82.2
CLIP→MIM [34, 44, 63]	–	66.2	82.4
RILS (E2E)	**37.5**	**68.5**	**82.7**

Table 9. **Comparisons with two-stage methods.** All methods are trained with ViT-B/16 on L-10M for 25 epochs. Our approach rivals all two-stage methods in terms of zero-shot (ZS.), linear probing (Lin.) and end-to-end fine-tuning (FT.)

Reconstruction Space	ZS.	Lin.	FT.
Raw Pixel Space (MAE+CLIP)	34.2	61.9	82.2
High-level Vision Space [12, 75]	34.8	67.7	82.4
Language Semantic Space (RILS)	**37.5**	**68.5**	**82.7**

Table 10. **Comparisons on reconstruction space.** Raw pixel space denotes aforementioned MAE+CLIP which directly reconstruct RGB values. High-level vision space indicates to replace the language feature z^T in Eq. (8) to a group of learnable embeddings similar to [1, 7, 75]. Language semantic space stands for ours. All methods are trained with ViT-B/16 for 25 epochs on L-10M.

RILS shows better performance on all three metrics. We notice that compared to reconstruct raw pixels, the high-level space reconstruction does not shown prominent advantages as in [1, 7, 75]. We guess this is due to the absence of multi-crop augmentation and exponential moving average (EMA) strategy which play important roles inside these methods but bring significant adverse impact on training efficiency.

5.3. Hyper-parameters Analysis

Mask Ratio. We compare three different mask ratios in Table 11a. As results shown, randomly masking at a ratio of 75% leads to generally decent performance. Lower mask ratio (60%) leads to +0.8% better linear probing accuracy but impairs zero-shot performance (−1.6%). Higher mask ratio (90%) acquires similar zero-shot accuracy but results in apparent decrease of −0.9% and −0.4% in terms of linear probing and fine-tuning. We guess this is due to lower mask ratio brings more prior information but less supervision in reconstruction, while the higher one exact converses.

Number of vision decoder blocks. Performance w.r.t. number of decoder blocks are shown in Table 11b. Different decoder blocks numbers exhibit less to none differences in terms of zero-shot. As to linear probing, we detect the same trend as [28]: More vision decoder blocks leads to better linear probing accuracy. For fine-tuning, we observe almost the same accuracy but more decoder blocks seems to be slightly worse. Besides the numerical differences, when take the extra costs bring by more decoder blocks into consideration, we choose to build our RILS with one vision decoder blocks only, due to more decoder blocks introduces

Mask Ratio	ZS.	Lin.	FT.
60%	35.9	69.3	82.7
75%	37.5	68.5	82.7
90%	37.4	67.6	82.3

(a) **Ablations on mask ratio.** 75% generally leads to a good result except for linear probing.

Nums.	ZS.	Lin.	FT.	Rel.GHs.
1	37.5	68.5	82.7	1.0×
2	37.8	68.9	82.6	1.3×
4	37.3	69.6	82.5	1.5×

(b) **Numbers of vision decoder blocks.** Rel.GHs. is short for relative GPU hours.

λ_1:λ_2	ZS.	Lin.	FT.
2:1	37.5	68.5	82.7
1:1	36.4	68.1	82.3
1:2	35.9	68.2	82.2

(c) **Loss coefficients** of $\mathcal{L}_{\text{Contra}}$ and $\mathcal{L}_{\text{Recon}}$.

Table 11. Ablations on the hyper-parameters of our approach. We report zero-shot (ZS.), linear probing(Lin.) and end-to-end fine-tuning (FT.) accuracy (%) on ImageNet-1K. Our default setups are high-lighted in gray.

un-negligible training overheads. For instance, increasing the number of decoder blocks from 1 to 4 brings 0.5× more training costs.

Loss coefficients. The results under different loss coefficients in Eq. (6) are shown in Table 11c. We empirically observe better performance with λ_1:λ_2 setting to 2:1.

6. Conclusion

In this work, we present a unified vision representation learner RILS that subsumes masked image modeling with natural language supervision. Instead of simply combining both paradigms with limited connection, we present a novel design to perform MIM in the language semantic space. Text features from language encoder serves as basic prototypes and probabilistic distribution of masked patches explicitly . Both objectives complement each other, leading to high synergy and mutual benefits. Extensive experimental results on downstream tasks showcase our method's advanced transferability and out-of-the-box capacity. We also observe excellent properties emerge from our design, especially in the low-shot regime. We hope our work can provide a new perspective on how to utilize language supervision with masked image modeling. In the future, we will explore further scale-up of our approach in terms of both model size and data size.

Acknowledgement

This work was supported by NSFC (No. 62276108).

References

[1] Mahmoud Assran, Mathilde Caron, Ishan Misra, Piotr Bojanowski, Florian Bordes, Pascal Vincent, Armand Joulin, Mike Rabbat, and Nicolas Ballas. Masked siamese networks for label-efficient learning. In *ECCV*, 2022. 6, 7, 8

[2] Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. Data2vec: A general framework for self-supervised learning in speech, vision and language. *arXiv preprint arXiv:2202.03555*, 2022. 2

[3] Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021. 1, 2, 4, 5, 6

[4] Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. *NeurIPS*, 2019. 7

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *NeurIPS*, 2020. 1

[6] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *NeurIPS*, 2020. 6

[7] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021. 5, 7, 8

[8] 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 *CVPR*, 2021. 2

[9] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pre-training from pixels. In *ICML*, 2020. 2

[10] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020. 4

[11] Xiaokang Chen, Mingyu Ding, Xiaodi Wang, Ying Xin, Shentong Mo, Yunhao Wang, Shumin Han, Ping Luo, Gang Zeng, and Jingdong Wang. Context autoencoder for self-supervised representation learning. *arXiv preprint arXiv:2202.03026*, 2022. 2

[12] Yabo Chen, Yuchen Liu, Dongsheng Jiang, Xiaopeng Zhang, Wenrui Dai, Hongkai Xiong, and Qi Tian. Sdae: Self-distillated masked autoencoder. In *ECCV*, 2022. 2, 3, 8

[13] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *CVPRW*, 2020. 13

[14] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005. 2

[15] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 2, 6

[16] Karan Desai and Justin Johnson. Virtex: Learning visual representations from textual annotations. In *CVPR*, 2021. 2

[17] Karan Desai, Gaurav Kaul, Zubin Aysola, and Justin Johnson. Redcaps: Web-curated image-text data created by the people, for the people. *arXiv preprint arXiv:2111.11431*, 2021. 2

[18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 1, 2

[19] Xiaoyi Dong, Jianmin Bao, Ting Zhang, Dongdong Chen, Weiming Zhang, Lu Yuan, Dong Chen, Fang Wen, and Nenghai Yu. Peco: Perceptual codebook for bert pre-training of vision transformers. *arXiv preprint arXiv:2111.12710*, 2021. 2

[20] Xiaoyi Dong, Yinglin Zheng, Jianmin Bao, Ting Zhang, Dongdong Chen, Hao Yang, Ming Zeng, Weiming Zhang, Lu Yuan, Dong Chen, et al. Maskclip: Masked self-distillation advances contrastive language-image pretraining. *arXiv preprint arXiv:2208.12262*, 2022. 4

[21] 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. 1, 2, 12

[22] Yuxin Fang, Shusheng Yang, Shijie Wang, Yixiao Ge, Ying Shan, and Xinggang Wang. Unleashing vanilla vision transformer with masked image modeling for object detection. *arXiv preprint arXiv:2204.02964*, 2022. 1

[23] Philip Gage. A new algorithm for data compression. *C Users Journal*, 1994. 4

[24] Golnaz Ghiasi, Yin Cui, Aravind Srinivas, Rui Qian, Tsung-Yi Lin, Ekin D Cubuk, Quoc V Le, and Barret Zoph. Simple copy-paste is a strong data augmentation method for instance segmentation. In *CVPR*, 2021. 13

[25] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *NeurIPS*, 2020. 6

[26] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*, 2021. 1

[27] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *CVPR*, 2019. 5, 13

[28] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022. 1, 2, 4, 5, 6, 7, 8, 12

[29] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020. 2

[30] Kaiming He, Ross Girshick, and Piotr Dollár. Rethinking imagenet pre-training. In *ICCV*, 2019. 6

[31] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017. 5

[32] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *ICCV*, 2021. 7

[33] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *CVPR*, 2021. 7

[34] Zejiang Hou, Fei Sun, Yen-Kuang Chen, Yuan Xie, and Sun-Yuan Kung. Milan: Masked image pretraining on language assisted representation. *arXiv preprint arXiv:2208.06049*, 2022. 2, 3, 4, 7, 8

[35] 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 *ICML*, 2021. 1, 2

[36] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 4, 12, 13

[37] Xiaotong Li, Yixiao Ge, Kun Yi, Zixuan Hu, Ying Shan, and Ling-Yu Duan. mc-beit: Multi-choice discretization for image bert pre-training. In *ECCV*, 2022. 2

[38] 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. 2, 12

[39] Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. *arXiv preprint arXiv:2203.16527*, 2022. 1, 5

[40] 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 *ECCV*, 2014. 5, 6, 13

[41] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. In *CVPR*, 2022. 2

[42] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021. 1, 2

[43] Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. Slip: Self-supervision meets language-image pre-training. In *European Conference on Computer Vision*, 2022. 2, 4, 5, 6, 7, 12

[44] Zhiliang Peng, Li Dong, Hangbo Bao, Qixiang Ye, and Furu Wei. Beit v2: Masked image modeling with vector-quantized visual tokenizers. *arXiv preprint arXiv:2208.06366*, 2022. 2, 4, 7, 8

[45] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *NAACL*, 2018. 1

[46] 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 *ICCV*, 2015. 6

[47] 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 *ICML*, 2021. 1, 2, 4, 5, 6, 7, 8, 12

[48] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. *OpenAI blog*, 2018. 1

[49] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 2019. 1

[50] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 1

[51] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *ICML*, 2021. 2

[52] Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. Denseclip: Language-guided dense prediction with context-aware prompting. In *CVPR*, 2022. 1

[53] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *ICML*, 2019. 7

[54] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *NeurIPS*, 2015. 5

[55] Jason Tyler Rolfe. Discrete variational autoencoders. *arXiv preprint arXiv:1609.02200*, 2016. 2

[56] Mert Bulent Sariyildiz, Julien Perez, and Diane Larlus. Learning visual representations with caption annotations. In *ECCV*, 2020. 2

[57] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2210.08402*, 2022. 2

[58] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 2, 4

[59] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, 2018. 2

[60] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 2017. 1, 2, 3

[61] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. *NeurIPS*, 2019. 7

[62] Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. In *CVPR*, 2022. 2

[63] Longhui Wei, Lingxi Xie, Wengang Zhou, Houqiang Li, and Qi Tian. Mvp: Multimodality-guided visual pre-training. *arXiv preprint arXiv:2203.05175*, 2022. 2, 3, 4, 7, 8

[64] Bichen Wu, Ruizhe Cheng, Peizhao Zhang, Peter Vajda, and Joseph E Gonzalez. Data efficient language-supervised zero-shot recognition with optimal transport distillation. *arXiv preprint arXiv:2112.09445*, 2021. 2

[65] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *ECCV*, 2018. 5

[66] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *CVPR*, 2022. 2

[67] Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Bin Xiao, Ce Liu, Lu Yuan, and Jianfeng Gao. Unified contrastive learning in image-text-label space. In *CVPR*, 2022. 2

[68] 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. *arXiv preprint arXiv:2111.07783*, 2021. 2

[69] Haoxuan You, Luowei Zhou, Bin Xiao, Noel Codella, Yu Cheng, Ruochen Xu, Shih-Fu Chang, and Lu Yuan. Learning visual representation from modality-shared contrastive language-image pre-training. In *European Conference on Computer Vision*. Springer, 2022. 2, 4

[70] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022. 2

[71] 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 *CVPR*, 2022. 7, 8

[72] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022. 1

[73] Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. Contrastive learning of medical visual representations from paired images and text. *arXiv preprint arXiv:2010.00747*, 2020. 2

[74] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 2019. 5

[75] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. *arXiv preprint arXiv:2111.07832*, 2021. 2, 3, 7, 8