Towards All-in-one Pre-training via Maximizing Multi-modal Mutual Information

Weijie Su[1*†], Xizhou Zhu[2,4*✉], Chenxin Tao[3*†], Lewei Lu[2], Bin Li[1], Gao Huang[3],
Yu Qiao[4], Xiaogang Wang[5,2], Jie Zhou[3], Jifeng Dai[3,4]
[1]University of Science and Technology of China [2]SenseTime Research [3]Tsinghua University
[4]Shanghai Artificial Intelligence Laboratory [5]The Chinese University of Hong Kong
jackroos@mail.ustc.edu.cn, {zhuwalter,luotto}@sensetime.com
tcx20@mails.tsinghua.edu.cn, binli@ustc.edu.cn, {gaohuang,jzhou,daijifeng}@tsinghua.edu.cn
qiaoyu@pjlab.org.cn, xgwang@ee.cuhk.edu.hk

Abstract

*To effectively exploit the potential of large-scale models, various pre-training strategies supported by massive data from different sources are proposed, including supervised pre-training, weakly-supervised pre-training, and self-supervised pre-training. It has been proved that combining multiple pre-training strategies and data from various modalities/sources can greatly boost the training of large-scale models. However, current works adopt a multi-stage pre-training system, where the complex pipeline may increase the uncertainty and instability of the pre-training. It is thus desirable that these strategies can be integrated in a single-stage manner. In this paper, we first propose a general multi-modal mutual information formula as a unified optimization target and demonstrate that all mainstream approaches are special cases of our framework. Under this unified perspective, we propose an all-in-one single-stage pre-training approach, named **M**aximizing **M**ulti-modal **M**utual **I**nformation Pre-training (**M3I Pre-training**). Our approach achieves better performance than previous pre-training methods on various vision benchmarks, including ImageNet classification, COCO object detection, LVIS long-tailed object detection, and ADE20k semantic segmentation. Notably, we successfully pre-train a billion-level parameter image backbone and achieve state-of-the-art performance on various benchmarks under public data setting. Code shall be released at https://github.com/OpenGVLab/M3I-Pretraining.*

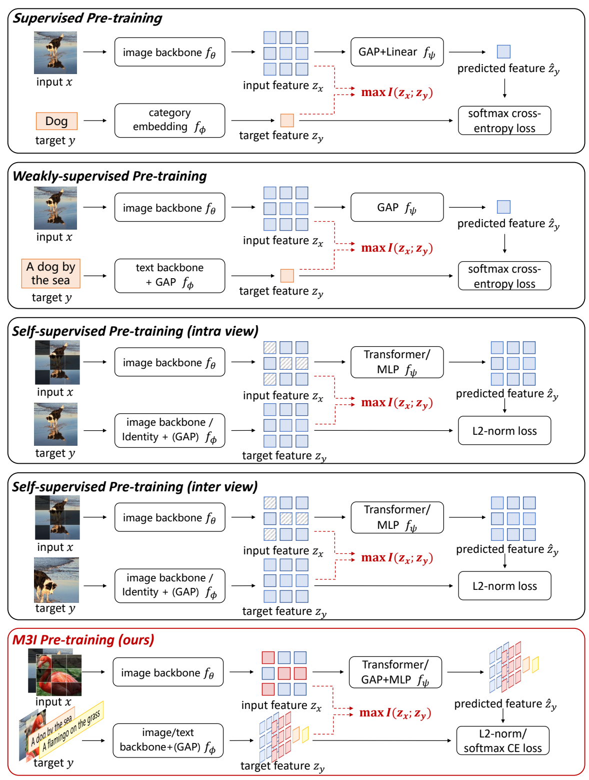

Figure 1. Comparison between different pre-training paradigms and M3I Pre-training. Existing pre-training methods are all optimizing the mutual information between the input and target representations, which can be integrated by M3I Pre-training. "GAP" refers to global average pooling.

1. Introduction

In recent years, large-scale pre-trained models [5, 13, 27, 30, 37, 55, 65, 91] have swept a variety of computer vision tasks with their strong performance. To adequately train large models with billions of parameters, researchers design various annotation-free self-training tasks and obtain sufficiently large amounts of data from various modalities and sources. In general, existing large-scale pre-training strategies are mainly divided into three types: supervised learn-

* Equal contribution. †This work is done when Weijie Su and Chenxin Tao are interns at Shanghai Artificial Intelligence Laboratory. ✉ Corresponding author.

ing [20, 65] on pseudo-labeled data (*e.g.,* JFT-300M [85]), weakly supervised learning [37,55] on web crawling images text pairs (*e.g.,* LAION-400M [56]), and self-supervised learning [5, 13, 27, 30, 91] on unlabeled images. Supported by massive data, all these strategies have their own advantages and have been proven to be effective for large models of different tasks. In pursuit of stronger representations of large models, some recent approaches [47, 78, 81] combine the advantages of these strategies by directly using different proxy tasks at different stages, significantly pushing the performance boundaries of various vision tasks.

Nevertheless, the pipeline of these multi-stage pre-training approaches is complex and fragile, which may lead to uncertainty and catastrophic forgetting issues. Specifically, the final performance is only available after completing the entire multi-stage pre-training pipeline. Due to the lack of effective training monitors in the intermediate stages, it is difficult to locate the problematic training stage when the final performance is poor. To eliminate this dilemma, it is urgent to develop a single-stage pre-training framework that can take advantage of various supervision signals. It is natural to raise the following question: *Is it possible to design an all-in-one pre-training method to have all the desired representational properties?*

To this end, we first point out that different single-stage pre-training methods share a unified design principle through a generic pre-training theoretical framework. We further extend this framework to a multi-input multi-target setting so that different pre-training methods can be integrated systematically. In this way, we propose a novel single-stage pre-training method, termed M3I Pre-training, that all desired representational properties are combined in a unified framework and trained together in a single stage.

Specifically, we first introduce a generic pre-training theoretical framework that can be instantiated to cover existing mainstream pre-training methods. This framework aims to maximize the mutual information between input representation and target representation, which can be further derived into a prediction term with a regularization term. (1) The prediction term reconstructs training targets from the network inputs, which is equivalent to existing well-known pre-training losses by choosing proper forms for the predicted distribution. (2) The regularization term requires the distribution of the target to maintain high entropy to prevent collapse, which is usually implemented implicitly through negative samples or stop-gradient operation. As shown in Fig. 1, by adopting different forms of input-target paired data and their representations, our framework can include existing pre-training approaches and provide possible directions to design an all-in-one pre-training method.

To meet the requirement of large-scale pre-training with various data sources, we further extend our framework to the multi-input multi-target setting, with which we show that multi-task pre-training methods are optimizing a lower bound of the mutual information. In addition, we mix two masked views from two different images as the input. The representation of one image is used to reconstruct the same view, while the other image is used to reconstruct a different augmented view. Both representations will predict their corresponding annotated category or paired texts. In this way, we propose a novel pre-training approach, called M3I Pre-training, which can effectively combine the merits of supervised/weakly-supervised/self-supervised pre-training and enables large-scale vision foundation models to benefit from multi-modal/source large-scale data. Our contributions can be summarized as follows:

- We theoretically demonstrate all existing mainstream pre-training methods share a common optimization objective, *i.e.,* maximizing the mutual information between input and target representation. We also show how to instantiate our framework as distinct pre-training methods.

- We propose a novel single-stage pre-training approach called M3I Pre-training to gather the benefit of various pre-training supervision signals, via extending our mutual information pre-training framework to a multi-input multi-target setting.

- Comprehensive experiments demonstrate the effectiveness of our approach. We successfully pre-train InternImage-H [79], a model with billion-level parameters, and set a new record on basic detection and segmentation tasks, *i.e.,* 65.4 box AP on COCO test-dev [46], 62.9 mIoU on ADE20K [98].

2. Related Work

Supervised Pre-training (SP) has been the mainstream over a long period of time [9, 11, 24, 26, 32, 48, 49]. Most works adopt image classification on ImageNet [21] as the pre-training task, for both ConvNets [32, 49, 65, 86] and Transformers [24, 48, 72]. SP has benefited many downstream tasks, including object detection [9,26], semantic semantation [11, 83], and video recognition [8]. Some works have also explored the scaling properties of pre-training datasets [24, 92] and image backbones [1, 92]. Moreover, SP shows that mixing two inputs [90, 94] is critical for improving accuracy [72, 73], which is rarely explored in other pre-training paradigms. Our proposed framework includes SP as a special case. M3I Pre-training can thus preserve the advantage of it in an all-in-one pre-training and surpass its performances on downstream tasks.

Self-supervised Pre-training (SSP) becomes popular in recent years [36, 93]. It does not require annotations, and thus enables the usage of large-scale unlabeled data. SSP can be divided into two kinds of methods: Intra-view tasks create input and target from the same view, which includes auto-encoder [34, 76], global/dense distillation [33, 81] and masked image modeling (MIM) [4, 5, 14, 30, 87]. On the contrary, inter-view tasks adopt different augmented views as input and target, such as dense/global instance discrimination [53, 82] and siamese image modeling [67]. Instance discrimination contains several sub-frameworks, including contrastive learning [13,31,80,84], asymmetric networks [15, 27] and feature decorrelation [6, 91], which are found of similar mechanism [66, 70]. Some SSP methods have displayed great potential by surpassing SP on downstream tasks by a large margin [13, 30, 31]. MIM has also been proven to work well with large-scale networks [30]. SiameseIM [67] is a inter-view SSP method that better combines semantic alignment with spatial sensitivity. Our method covers different self-supervised pre-training methods in a general framework and seeks to find the most effective setting through extensive experiments. It can combine all the strengths and shows impressive performances.

Weakly-supervised Pre-training (WSP) utilizes image-text datasets [10, 56, 58, 68] or image-hashtag datasets [51, 60, 75]. These pre-training methods rely on noisy supervision from the Internet and are thus scalable. For image-hashtag datasets, some works [51, 60] show competitive performances in various transfer-learning settings. For image-text datasets, earlier works [2, 16, 41, 42, 50, 61–64] mainly focused on learning general representations for visual-linguistic understanding. Recently CLIP [55] and ALIGN [37] demonstrated the effectiveness of image-text pre-training in image recognition. They propose to learn the aligned visual-linguistic representations and achieve impressive image classification accuracy in a zero-shot manner. In our framework, WSP is shown to be a special instance. M3I Pre-training leverages the power of WSP to achieve a new height on various downstream tasks.

Multi-task Pre-training adopts multiple targets for the same input. This kind of method usually combines text target from WSP and image target from SSP [23, 52, 59, 89]. Some works have also explored using both category targets from SP and image targets from SSP [39, 44]. Another works [77, 96] aim at unifying visual-linguistic understanding tasks and image perception tasks. Multi-task pre-training fits well into our framework with the multi-input multi-target extension. Compared with previous works, our method can be successfully applied to large-scale models and displays superior results.

Multi-stage Pre-training instead adopts stage-wise pre-training, which focuses on one pre-training target in each stage and reuses the model in the next stage [54, 78, 81]. Multi-stage pre-training also follows the mutual information objective in each stage. However, multi-stage pre-training suffers from a complex pipeline that may increase uncertainty and instability. On the contrary, M3I Pre-training combines different supervision signals in a single stage that avoids the problems of multi-stage pre-training.

3. Method

3.1. Mutual Information for Generic Pre-training

The goal of pre-training is to learn representations that can well represent the training samples. Suppose the training sample is s, where s could be image-category pair (supervised), image-text pair (weakly-supervised), or image only (self-supervised). The input training data x and target training data y are extracted from s by some transform operations (t_x, t_y), *i.e.,* $x = t_x(s), y = t_y(s)$. The transform operations t is typically data-irrelevant, *e.g.,* “apply image augmentation”, “get annotated category” or “get paired text”. In vision-centric pre-training, the input data x is usually an augmented image, while the target data y is either the annotated category, the paired text, or also an augmented image. z_x and z_y denote the encoded representation for input and target, respectively. Then, the desired pre-training objective can be described as maximizing the conditional mutual information between z_x and z_y given t_x and t_y as (see Appendix for derivation):

$$\begin{aligned}
&(s, t_x, t_y) \sim D_{\text{train}} && \text{(sampled training sample)},\\
&x = t_x(s)\ ,\ y = t_y(s) && \text{(extracted training data)},\\
&z_x \sim p(z_x|x)\ ,\ z_y \sim p(z_y|y) && \text{(encoded training representation)},
\end{aligned}$$

$$\begin{aligned}
I(z_x; z_y \mid t_x, t_y) = &\underbrace{\mathbb{E}_{p(t_y)}\Big[H\big(p(z_y|t_y)\big)\Big]}_{\text{regularization term to avoid collapse}} \\
&- \underbrace{\mathbb{E}_{p(s,t_x,t_y,z_x)}\Big[H\big(p(z_y|y)\ ,\ p(z_y|z_x,t_x,t_y)\big)\Big]}_{\text{(cross-entropy) prediction term for target representation}}, \quad (1)
\end{aligned}$$

where the first term requires high entropy of the target representation, which avoids collapse. The second term requires the posterior distribution $p(z_y|z_x, t_x, t_y)$ to be close to the target distribution $p(z_y|y)$.

In practice, deterministic neural networks are used for encoding the input representation z_x and target representation z_y. On the other hand, the posterior distribution $p(z_y|z_x, t_x, t_y)$ is usually intractable. To alleviate this issue, a common practice is introducing another parameterized distribution $p_\psi(z_y|z_x, t_x, t_y)$ as an approximation, Then, Eq. (1) becomes (see Appendix for derivation):

$$\begin{aligned}
&z_x = f_\theta(x), \quad z_y = f_\phi(y) && \text{(parameterized representation)},\\
&\hat{z}_y = f_\psi(z_x, t_x, t_y) && \text{(prediction of } z_y \text{ given } z_x),\\
&p_\psi(z_y|z_x, t_x, t_y) = \hat{P}(z_y \mid \hat{z}_y) && \text{(approximated distribution)},
\end{aligned}$$

$$\begin{aligned}
I(z_x; z_y \mid t_x, t_y) = \sup_{f_\psi} &\underbrace{\mathbb{E}_{p(t_y)}\Big[H\big(p(z_y(\phi) \mid t_y)\big)\Big]}_{\text{regularization term to avoid collapse}} \\
&+ \underbrace{\mathop{\mathbb{E}}_{p(s,t_x,t_y)}\Big[\log \hat{P}\big(z_y(\phi) \mid \hat{z}_y(\theta,\psi)\big)\Big]}_{\text{(log-likelihood) prediction term for target representation}}, \quad (2)
\end{aligned}$$

where $\hat{P}$ is the approximated posterior distribution of z_y given the prediction of $\hat{z}_y$. θ, ϕ, and ψ are learnable parameters of input encoder, target encoder, and input-to-target decoder, respectively. When z_y is continuous and deterministic given y, the regularization term becomes intractable [7, 71]. Different mechanisms would be incorporated to avoid representation collapse (see Sec 3.2). Then, to maximize the mutual information in Eq. (2), the training loss can be derived as:

$$\min_{\theta,\phi,\psi} \mathop{\mathbb{E}}_{p(s,t_x,t_y)} L(s, t_x, t_y; \theta, \phi, \psi) = -\log \hat{P}\big(z_y(\phi) \mid \hat{z}_y(\theta, \psi)\big),$$
$$\textit{s.t.} \quad \text{non-collapse representation of } z_y. \tag{3}$$

Different form of $\hat{P}$ results in different loss, *e.g.,* Gaussian and Boltzmann distributions corresponding to L2-norm and Softmax cross-entropy losses, respectively:

$$\hat{P}(z_y \mid \hat{z}_y) \sim \mathcal{N}(\hat{z}_y, \sigma^2 I) \quad \text{(Gaussian distribution)}$$
$$\Rightarrow L = -\log \hat{P}(z_y \mid \hat{z}_y) = \frac{1}{2\sigma^2}\|z_y - \hat{z}_y\|^2 + C,$$
$$\hat{P}(z_y \mid \hat{z}_y) \propto \exp(\hat{z}_y^T z_y/\tau) \quad \text{(Boltzmann distribution)}$$
$$\Rightarrow L = -\log \hat{P}(z_y \mid \hat{z}_y) = -\log \frac{\exp(\hat{z}_y^T z_y/\tau)}{\sum_{z'_y} \exp(\hat{z}_y^T z'_y/\tau)},$$

where σ and τ are the hyper-parameters of Gaussian and Boltzmann distributions, respectively. C is a constant that can be ignored. z'_y iterates over all possible target representations $\{f_\phi(y) \mid y = t_y(s) \in D_{\text{train}}\}$.

Eq. (3) is a generic pre-training loss that can be instantiated into different pre-training paradigms, including supervised, weakly-supervised, and self-supervised pre-training. Tab. 5 in Appendix demonstrate the actual implementation of mainstream pre-training methods, which incorporates different mechanisms to avoid representation collapse.

3.2. Connection with Existing Pre-training Methods

Supervised Pre-training (SP) usually adopts *Image Classification (IC)* as the pre-training task. It takes an augmented image I as input data and the corresponding annotated category C as the target data. The input representation is $z_x = f_\theta(I)$, while the target representation is the category embedding (*e.g.,* linear classification weight) $z_y = f_\phi(C)$. The classifier predicts the category based on z_x as $\hat{z}_y = f_\psi(z_x)$. Thus, the pre-training objective is to maximize $I\big(f_\theta(I)\,;\,f_\phi(C)\big)$, and the SP loss can be derived as minimizing $L = -\log \hat{P}\big(f_\phi(C) \mid f_\psi \circ f_\theta(I)\big)$.

$$\max\ I\big(f_\theta(I)\,;\,f_\phi(C)\big)$$
$$\Rightarrow \min\ L = -\log \hat{P}\big(f_\phi(C) \mid f_\psi \circ f_\theta(I)\big),$$

where $\hat{P}$ is typically Boltzmann distribution (*i.e.,* Softmax cross-entropy loss). This distribution contains negative categories and naturally prevents collapse. As a mainstream pre-training framework, SP has been proven to be helpful on many downstream tasks over a long period of time [9, 11, 26, 83]. It learns from clean human-annotated data. This helps the model to develop common semantics and converge faster on downstream tasks.

Weakly-supervised Pre-training (WSP) usually adopts *Contrastive Language-Image Pre-training (CLIP)* [37, 55] as the pre-training task. It takes an augmented image I as input, and the corresponding paired text T as targets. Similar to supervised learning, the pre-training objective is

$$\max\ I\big(f_\theta(I)\,;\,f_\phi(T)\big)$$
$$\Rightarrow \min\ L = -\log \hat{P}\big(f_\phi(T) \mid f_\psi \circ f_\theta(I)\big),$$

where $\hat{P}$ is also Boltzmann distribution, which contains negative samples to prevent the collapse. WSP is able to exploit the massive image-text pairs from the Internet. With the help of image-text alignment, it not only enables many possible new tasks, *e.g.,* open-vocabulary recognition [28, 55], but also greatly boosts the performances of classification and detection tasks in long-tail scenario [69].

Self-supervised Pre-training (SSP) learns representation using images only. Given a sampled training image I, the input data is an augmented view of this image $\tilde{I}_x = t_x(I)$, the target data is another augmented view $\tilde{I}_y = t_y(I)$. The pre-training objective is derived from Eq. (3) as

$$\max\ I\big(f_\theta(\tilde{I}_x)\,;\,f_\phi(\tilde{I}_y)\big)$$
$$\Rightarrow \min\ L = -\log \hat{P}\Big(f_\phi(\tilde{I}_y) \mid f_\psi\big(f_\theta(\tilde{I}_x), t_x, t_y\big)\Big),$$

where t_x and t_y are the input and target augmentations on the sampled image, respectively. Depending on different methods, the target encoder f_ϕ could be identity, shared with f_θ or the Exponential Moving Average (EMA) of f_θ. $\hat{P}$ is usually Boltzmann or Gaussian distribution. When $\hat{P}$ is Boltzmann (*i.e.,* Softmax cross-entropy loss), it aims to differentiate z_y from different training data. When $\hat{P}$ is Gaussian (*i.e.,* L2-norm loss), it fits the value of z_y. To prevent collapse, "stop-gradient" [27], feature-decorrelation [91] and negative samples [13] are considered.

As Tab. 5 illustrated, different choices of data transform operations (t_x, t_y) and target representation type z_y result in different pre-training tasks: (1) For (t_x, t_y), they could be either the the same view (*e.g.,* auto-encoder) or different views (*e.g.,* instance discrimination [13, 27, 91]). t_x could also incorporate an additional mask operation (*e.g.,* masked image modeling [5, 30]). (2) For z_y, its representation type could be from $\{$dense pixels, dense feature, global feature$\}$.

The advantage of SSP methods is that they can utilize large-scale unlabelled data, which facilitates the development of large models. Some SSP methods can already surpass SP on downstream tasks [5, 30, 31]. Notably, MIM [5, 30] demonstrates great dense localization ability, while SiameseIM [67] can exhibit semantic alignment and spatial sensitivity at the same time.

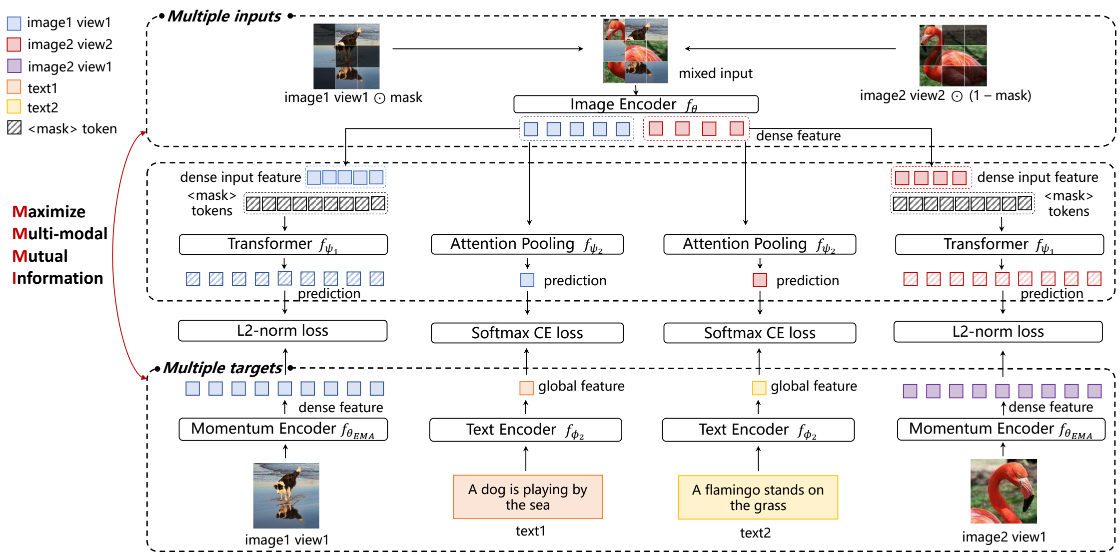

Figure 2. Overview of M3I Pre-training. We mix two views from different images as the inputs. The first image needs to predict the same image view, while the other image needs to predict a different augmented view. Both images need to predict their annotated category or paired text. When predicting image targets, we add position embeddings to decoder inputs so that “<mask>” tokens can be matched with target patches. Following [67], position embeddings are computed with respect to the left top origin of the input views.

3.3. Multi-input Multi-target Pre-training

Different pre-training tasks possess their own strengths. We would like to maintain all these properties in one pre-training approach. For this purpose, we extend our framework to a multi-input multi-target setting.

Suppose the set of N multiple inputs and M multiple targets are $X = \{x_i\}_{i=1}^N$ and $Y = \{y_j\}_{j=1}^M$, respectively. We use $\boldsymbol{t}_x$ and $\boldsymbol{t}_y$ to indicate the sets of transforms of inputs and targets. In practice, most methods choose to optimize the objectives of different types of targets separately. In this case, we can split the targets Y into K non-overlapping groups and encode different groups independently as $Y_m \cap Y_{n\neq m} = \varnothing, \cup_{k=1}^K Y_k = Y$. With this modification, we show that the mutual information in Eq. (2) can be bounded by (see Appendix for derivation):

$$
\begin{aligned}
&(\boldsymbol{s}, \boldsymbol{t}_x, \boldsymbol{t}_y, X, Y) \sim D_{\text{train}} && \text{(sample inputs and targets)} \\
&\boldsymbol{z}_x = f_\theta\big(X = \{x_i\}_{i=1}^N\big) && \text{(encode multiple inputs jointly)} \\
&\boldsymbol{z}_y^k = f_{\phi_k}(Y_k),\ Y_k = \{y_{k_j}\}_{j=1}^{M_k} && \text{(encode multiple targets separately)} \\
&\hat{\boldsymbol{z}}_y^k = f_{\psi_k}(\boldsymbol{z}_x, \boldsymbol{t}_x, \boldsymbol{t}_y) && \text{(predict multiple targets separately)}
\end{aligned}
$$

$$
\begin{aligned}
I\big(\boldsymbol{z}_x; \{\boldsymbol{z}_y^k\}_{k=1}^K | \boldsymbol{t}_x, \boldsymbol{t}_y\big) \geq &\sup_{\{f_{\psi_k}\}_{k=1}^K} \underbrace{\mathop{\mathbb{E}}_{p(\boldsymbol{t}_y)} \Big[H\Big(p\big(\{\boldsymbol{z}_y^k\}_{k=1}^K | \boldsymbol{t}_y\big)\Big)\Big]}_{\text{regularization term to avoid collapse}} \\
&+ \underbrace{\sum_{k=1}^K \mathop{\mathbb{E}}_{p(\boldsymbol{s}, \boldsymbol{t}_x, \boldsymbol{t}_y)} \Big[\log \hat{P}_k\big(\boldsymbol{z}_y^k \mid \hat{\boldsymbol{z}}_y^k\big)\Big]}_{\text{(log-likelihood) prediction term for target representation}}, \\
\Rightarrow L(\boldsymbol{s}, \boldsymbol{t}_x, \boldsymbol{t}_y) = &\sum_{k=1}^K -\log \hat{P}_k\big(\boldsymbol{z}_y^k(\phi_k) \mid \hat{\boldsymbol{z}}_y^k(\theta, \psi_k)\big),
\end{aligned} \tag{4}
$$

where k is the group index, M_k is the number of targets in k^{th} group, and $\hat{P}_k$ is the approximated distribution for each target group. Each prediction term corresponds to the objective of a target group. This implies that optimizing target objectives independently is equivalent to optimizing a lower bound of the mutual information.

Multi-input Pre-training ($N = M$) uses multiple inputs with one target for each input $X = \{x_i\}_{i=1}^N, Y = \{y_i\}_{i=1}^N$,

where y_i is the corresponding sampled target of x_i. multi-input pre-training is widely used in supervised pre-training (typically $N = 2$), where different images are mixed through Mixup [94] or CutMix [90]. It has proven to be critical for improving accuracy and providing a more stable generalization ability. However, in other pre-training paradigms, a single input is usually adopted. The lack of multiple inputs may hinder better model performance, and also lead to inconsistency between different pre-training paradigms, which hampers the pre-training combination.

Multi-target Pre-training ($N = 1$) only uses multiple targets for the same input as $X = x, Y = \{y_i\}_{i=1}^M$.

Some previous works have explored the use of multiple targets [39, 44, 59, 89]. One line of research tries to combine weakly-supervised pre-training with specific forms of self-supervised pre-training, such as MaskCLIP [23] and FLAVA [59]. Another line studies the combination of supervised pre-training and self-supervised pre-training, such as SupCon [39] or SupMAE [44]. These methods display the effectiveness of multiple targets.

3.4. M3I Pre-training

With the help of our mutual information framework, we are able to systematically integrate different pre-trainings into a whole, which we name as M3I Pre-training. It has two inputs and four targets, combining self-supervised and supervised / weakly-supervised pre-training as

$$X = \{\tilde{I}_{x_i}, \tilde{I}_{x_j}\}, \quad Y = \{\tilde{I}_{y_i}, \tilde{I}_{y_j}, T_i, T_j\},$$

where $\tilde{I}_{x_i}, \tilde{I}_{x_j}$ are the augmented input views of two different sampled images I_i, I_j (in implementation, for a mini-batch of size N, we randomly shuffle the indexes $\{1, 2, ..., N\}$ to $\{\sigma(1), \sigma(2), ..., \sigma(N)\}$ and set $j = \sigma(i)$). $\tilde{I}_{y_i}, \tilde{I}_{y_j}$ are the corresponding augmented target views. T_i, T_j denotes the corresponding annotated category (supervised) or paired text (weakly-supervised) for each image.

Input Encoder f_θ first mixes the input views with a randomized binary mask m as $\tilde{I}_{\text{mix}} = m \odot \tilde{I}_{x_i} + (1-m) \odot \tilde{I}_{x_j}$, where $\odot$ is the element-wise product, m shares the same shape as inputs. Then, the input representation is encoded by an image backbone (*e.g.,* ViT [24]) as $f_\theta(\tilde{I}_{\text{mix}})$. In order to make this mix strategy compatible with existing pre-training tasks like Masked Image Modeling (MIM) and Image Classification (IC), we split the mask m into patches with $p \times p$ size. All pixels in the same patch will be masked or unmasked together. For example, $p = 16$ is by default used for MIM [5, 30]. Note that the widely used Mixup [94] and CutMix [90] are generally incompatible with MIM.

Target Encoder f_ϕ is responsible for producing the target representations. For image targets $\tilde{I}_{y_i}, \tilde{I}_{y_j}$, we use the momentum input image backbone as the encoder to generate dense target features. Note that for a mini-batch of N samples, we only need to forward target encoder N times (one for each target image $\tilde{I}_{y_i}$). For category targets (supervised) or text targets (weakly-supervised) T_i, T_j, we use a category embedding or text backbone that is jointly trained during pre-training. Notice that because of the multiple inputs, we can adopt both intra-view and inter-view self-supervised predictions: the first image i is asked to predict the same view (*i.e.,* $\tilde{I}_{x_i} = \tilde{I}_{y_i}$), and the other image j instead needs to predict a different augmented view (*i.e.,* $\tilde{I}_{x_j} \neq \tilde{I}_{y_j}$).

Input-to-Target Decoder f_ψ predicts the target representations from the input. For simplicity, we use the separate loss form in Eq. (4) to predict each target separately. We adopt Transformer [74] layers to predict the dense representations for image targets, and an attention pooling layer [14] followed by a linear projection to predict the category embedding (supervised) or text embedding (weakly-supervised).

4. Experiment

Implementation Details. We utilize InternImage-H [79] as image encoder in Sec 4.1 for large-scale model pre-training and ViT-B/16 [24] as that in other experiments for ablation study and fair comparison. For image-text dataset (*e.g.,* YFCC-15M [68]), a 12-layer Transformer (with the same network architecture as BERT-Base [22]) is utilized as text target encoder. For image classification dataset (*e.g.,* ImageNet-1k [21]), we directly use the linear classifier weight as category embedding target. We employ 4-layer Transformer as decoder for image representation target, and Attention Pooling as that for category embedding or text global feature. Please see Appendix for detailed pre-training hyper-parameters.

4.1. Pre-training of 1B Image Backbone

Settings. We employ InternImage-H [79] (a ConvNet-based image backbone with 1B parameters) as image encoder. The network is pre-trained for 30 epochs on 427M public image-text pairs (LAION400M [56], YFCC-15M [38], CC12M [10]) and 15M public image-category pairs (ImageNet-22k [21]). We report the transfer performance on ImageNet [21], COCO [46], LVIS [29], and ADE20k [97] benchmarks.

Results and Discussions. As shown in Tab. 1, all previous large model pre-training approaches adopt a complicated multi-stage training pipeline. Instead, our M3I Pre-training is a simple yet effective single-stage pre-training paradigm. It achieves state-of-the-art performance on COCO object detection, LVIS long-tailed object detection, and ADE20k semantic segmentation. Very competitive performance is achieved on ImageNet classification. It validates the effectiveness of our approach. Besides, M3I Pre-training only employs pubic datasets and exhibits superior transfer performance while all other approaches include private datasets in their pre-training.

Different from SwinV2 [47], BEiT-3 [78] and FD [81], M3I Pre-training is an all-in-one single-stage training paradigm which brings the following advantages: 1) *Simplicity.* M3I Pre-training could make use of all available supervision signals and training data in a single-stage pre-training. In contrast, both [47, 78] incorporate redundant multi-stage pre-training pipelines. [47] uses the same training data in multiple pre-training stages but with different supervision signals. [78] picks the pre-trained model in the previous pre-training stage as the target network for the next pre-training stage. 2) *Avoiding Catastrophic Forgetting*. As shown in Tab. 1, [47, 78, 81] all consist of multiple pre-training stages. The networks are expected to learn different representational attributes in different pre-training stages. However, due to the existence of catastrophic forgetting [25], attributes learned in the previous pre-training stage may be forgotten in the next pre-training stage. Our M3I Pre-training naturally avoids the catastrophic forgetting issue by learning different representational attributes simultaneously in one-stage pre-training.

Pre-training Approach	Model	Pipeline	Public Data	Private Data	ImageNet val	COCO test-dev	LVIS minival	ADE20k val
M3I Pre-training	InternImage-H [79] (1B)	Single Stage: M3I Pre-training	427M image-text 15M image-category	-	89.2	**65.4**	**62.5**	**62.9**
[47]	SwinV2-G (3B)	Stage 1: Masked Image Modeling$_{\text{pixel}}$ Stage 2: Image Classification	15M image-category	55M image-category	89.2	63.1	-	59.9
[78]	BEiT-3 (2B)	Stage 1: CLIP Stage 2: Dense Distillation Stage 3: Masked Data Modeling	21M image-text 15M image-category	400M image-text	**89.6**	63.7	-	62.8
[81]	SwinV2-G (3B)	Stage 1: Masked Image Modeling$_{\text{pixel}}$ Stage 2: Image Classification Stage 3: Dense Distillation	15M image-category	55M image-category	89.4	64.2	-	61.4
†previous best					89.1^{a}	64.5^{b}	59.8^{c}	60.8^{d}

Table 1. Comparision of M3I Pre-training with existing large model pre-training methods for visual recognition. Top1 Accuracy, AP$^{\text{box}}$, AP$^{\text{box}}$, mIoU are reported on ImageNet validation set, COCO test-dev set, LVIS minival set (to avoid data contamination following [96]), and ADE20k validation set, respectively. We achieve state-of-the-art performance on object detection and semantic segmentation tasks. M3I Pre-training also demonstrate very competitive classification performance with only public datasets, while all other methods utilize large-scale private data (WIT-400M [55] is used in [78], ImageNet-22k-ext [47] is used in [47,81]), which is strong correlated with the task of image classification. †We also list previous best results on these tasks with only public training data for comparision. Results reference: ^{a}MOAT [88], ^{b}Group DETR v2 [12], ^{c}GLIPv2 [96], ^{d}Mask DINO [40].

Pre-training Method	Input Data x	Target Representation z_y	ImageNet Top1	COCO AP$^{\text{box}}$
Self-supervised Pre-training (intra-view)				
(a) Auto-Encoder	view1	dense pixels	77.5	0.0†
(b) Dense Distillation	view1	dense feature	78.8	32.4
(c) Global Distillation	view1	global feature	77.1	27.9
(d) Masked Image Modeling$_{\text{pixel}}$	masked view1	dense pixels	83.1	46.8
(e) Masked Image Modeling$_{\text{feat}}$	**masked view1**	**dense feature**	**83.3**	**47.4**
(f) Masked Image Modeling$_{\text{gloal}}$	masked view1	global feature	83.2	47.5
Self-supervised Pre-training (inter-view)				
(g) Novel View Synthesis	view2	dense pixels	78.8	33.0
(h) Dense Instance Discrimination	view2	dense feature	83.2	50.1
(i) Instance Discrimination	view2	global feature	83.0	46.4
(j) Siamese Image Modeling$_{\text{pixel}}$	masked view2	dense pixels	78.9	38.1
(k) Siamese Image Modeling$_{\text{feat}}$	**masked view2**	**dense feature**	**83.7**	**49.8**
(l) Instance Discrimination$_{\text{mask}}$	masked view2	global feature	82.9	46.2

Table 2. Ablation study on different self-supervised pre-training methods under our framework. † denotes no convergence in fine-tuning.

Pre-training Method	ImageNet	COCO	LVIS		ADE20k
	Top1	AP$^{\text{box}}$	AP$^{\text{box}}$	AP$^{\text{box}}_{\text{rare}}$	mIoU
Image Classification	81.8	46.6	33.0	25.5	45.1
Best Intra-view SSP	83.3	47.4	31.2	21.9	40.1
Best Inter-view SSP	83.7	49.8	35.2	26.9	47.7
Ours					
M3I Pre-training w/o mix	83.7	50.3	36.6	27.2	48.7
M3I Pre-training	**83.9**	**50.8**	**37.5**	**29.6**	**49.0**

Table 3. Ablation study of multi-input multi-target pre-training.

4.2. Ablation Study

Ablation Settings. We utilize ViT-B/16 as the image backbone for the ablation study. The pre-training schedule is set to 400 epochs on ImageNet-1k. Different pre-training methods are evaluated by the transfer performance on ImageNet-1k classification, COCO detection, LVIS detection, and ADE20k segmentation. The fine-tuning schedule is 100 epochs for ImageNet-1k. For other datasets, fine-tuning with 25 epochs is adopted.

Ablation on Self-supervised Pre-training (SSP). As Tab. 5 shows, the mutual information framework proposes 12 forms of SSP, some of which have not been explored as pre-training before. We compare these 12 types of SSP in Tab. 2. Based on the experiment results, We analyze three key factors in these approaches:

1) Masked Input or Full Input. Masked input is critical for both intra-view and inter-view pre-training. The performances of Tab. 2 (d-f,j-l) (masked) are always better or on par with Tab. 2 (a-c,g-i)(full). The comparison for intra-view pre-training is consistent with previous studies [30], implying that masking operation can greatly boost the model's performance. We observe that the gap for inter-view pre-training becomes smaller. The reason may be that predicting another view constitutes a more challenging task, and reduce the information redundancy to some extent.

2) Target Representation. The dense feature works best under almost all settings. Compared to the global feature target, the dense feature target enables the spatial discrimination ability of the network. Thus, as shown in Tab. 2 (k,l) and Tab. 2 (h,i), it achieves much better performance on COCO. On the other hand, compared to dense pixels, dense features represent the target in high-level semantic space and thus bring semantic alignment capacity. For example, Tab. 2 (k) surpasses Tab. 2 (j) by a large margin both in ImageNet (+4.8 points) and COCO (+11.7 points).

3) Intra-view or Inter-view. The choice of intra-view or inter-view pre-training depends on whether the input data is masked or not. If full input is adopted, inter-view generally performs better than intra-view, as shown in Tab. 2 (a-c, g-i). We conjecture that recovering the same view is too easy, and may not be suitable for pre-training. On the other hand, if masked input is employed, both intra-view and inter-view can find a setting with good performance, *e.g.,* Tab. 2 (e,k).

Task	Metric	ImageNet Pre-train				YFCC Pre-train	
		SSP (intra-view)	SSP (inter-view)	SP	M3I (ImageNet)	WSP	M3I (YFCC)
ImageNet w/o Fine-tuning	Top1 acc.	×	×	**83.8** (DeiT-III)	83.3	†37.6 (CLIP)	†39.1
ImageNet Linear Classification	Top1 acc.	79.5 (iBOT)	78.0 (SiameseIM)	**83.8** (DeiT-III)	**83.8**	66.5 (CLIP)	72.3
ImageNet Fine-tuning	Top1 acc.	**84.2** (data2vec)	84.1 (SiameseIM)	83.8 (DeiT-III)	**84.2**	80.5 (CLIP)	83.7
COCO	AP^{box}	51.6 (MAE)	52.1 (SiameseIM)	47.6 (Sup.)	**52.2**	-	51.9
LVIS	AP^{box}	40.1 (MAE)	40.5 (SiameseIM)	37.2 (Sup.)	40.6	-	**40.8**
	AP^{box}_{rare}	38.1 (MAE)	38.1 (SiameseIM)	-	38.2	-	**38.4**
ADE20k	mIoU	50.0 (iBOT)	51.1 (SiameseIM)	49.3 (DeiT-III)	**51.3**	-	**51.3**

Table 4. System-level comparison with SoTA supervised, weakly-supervised, self-supervised pre-training methods (*i.e.,* DeiT-III [73], CLIP [55], data2vec [4], MAE [30], iBOT [99]and SiameseIM [67]). The results of CLIP are obtained from [52] for a fair comparison on YFCC-15M. The results of "Sup." (refers to supervised pre-training) are obtained from [43]. All methods adopt ViT-B/16 as the image backbone for a fair comparison. The fine-tuning schedule is 100 epochs for ImageNet, COCO, LVIS, and ADE20k. × denotes do not support. † "ImageNet w/o Fine-tuning" for weakly-supervised pre-training corresponds to the zero-shot transfer setting.

Ablation on Multi-input Multi-target Pre-training. After we have determined the best training setting for SSP with intra-view and inter-view, we can now combine different pre-training methods into an all-in-one approach with our multi-input multi-target framework. Tab. 3 demonstrates the comparison between M3I Pre-training and single-input single-target pre-training methods. Here we only consider pre-training on supervised pre-training for simplicity.

We first compare multi-target pre-training, *i.e.,* M3I Pre-training w/o mix[1], with single-input single-target pre-training methods. It's shown that M3I Pre-training w/o mix can obtain superior or comparable results on all tasks, especially on LVIS (+1.4 points) and ADE20k (+1.0 points) benchmarks. We note that even though some pre-training methods may not perform well on some tasks, the combination is still effective to improve upon all single-input single-target methods. This is because different pre-training methods focus on different representational properties. For example, Image Classification pre-training brings better semantic information. This leads to high results on LVIS and ADE20k datasets, where long-tail classes pose high demand for semantic understanding. Intra-view SSP instead excels on spatial sensitivity and delivers good performance on the COCO dataset. M3I Pre-training w/o mix demonstrates the benefits of these methods. Our final M3I Pre-training further adopts multiple inputs to better combine these pre-training methods. Experiments show that it achieves better performances on all tasks.

4.3. System-level Comparision with Other Methods

We compare M3I Pre-training with previous methods using the same ViT-B/16 [24] image backbone in Tab. 4. We pre-train our model for 1600 epochs and finetune it for 100 epochs on ImageNet [21], COCO [46], LVIS [29] and ADE20k [97] datasets. We also report the results on ImageNet without finetuning and with the linear protocol. We further validate our method on YFCC-15M image-text dataset [38]. For a fair comparison, the pre-training iterations are kept the same with ImageNet pre-training.

Tab. 4 shows that different pre-training methods possess different advantages. SP learns semantic alignment well and can already deliver good performance on ImageNet without further finetuning. WSP can enable zero-shot transfer learning, which can not be achieved through other pre-training methods. SSP presents better localization ability and is vital for dense prediction tasks. M3I Pre-training is able to achieve comparable results with the best of previous methods on all these tasks. This indicates that M3I Pre-training can maintain all these desired properties through a single-stage pre-training.

5. Conclusion

Modern large-scale networks rely on combining different pre-training methods to effectively utilize massive data, which suffers from the multi-stage pre-training practice. To derive a single-stage pre-training, we proposed a generic pre-training framework that unifies mainstream pre-training approaches. We further extended the framework to a multi-input multi-target setting, which shows that previous multi-task pre-training methods are actually optimizing a lower bound of the mutual information. Finally, we proposed an all-in-one pre-training method, M3I Pre-training. M3I Pre-training surpasses previous pre-training methods in various transfer-learning settings.

Limitations. We focused on vision-centric pre-training. The proposed framework can be applied to other domains, like natural language processing or visual-linguistic tasks. We expect to explore other domains in future work.

Acknowledgments The work is partially supported by the National Natural Science Foundation of China under grants No.U19B2044, No.61836011 and No.62022048. This work is also partially supported by the National Key R&D Program of China(NO.2022ZD0160100), and in part by Shanghai Committee of Science and Technology (Grant No. 21DZ1100100).

[1] "w/o mix" combines Inter-view SSP with Image Classification

References

[1] Samira Abnar, Mostafa Dehghani, Behnam Neyshabur, and Hanie Sedghi. Exploring the limits of large scale pre-training. *arXiv preprint arXiv:2110.02095*, 2021. 2

[2] Chris Alberti, Jeffrey Ling, Michael Collins, and David Reitter. Fusion of detected objects in text for visual question answering. *arXiv preprint arXiv:1908.05054*, 2019. 3

[3] 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*, pages 456–473. Springer, 2022. 14

[4] 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. 3, 8, 14

[5] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. In *ICLR*, 2021. 1, 2, 3, 4, 6, 14, 17

[6] Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv preprint arXiv:2105.04906*, 2021. 3

[7] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and R Devon Hjelm. Mine: mutual information neural estimation. *arXiv preprint arXiv:1801.04062*, 2018. 4

[8] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, 2021. 2

[9] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, pages 213–229. Springer, 2020. 2, 4

[10] 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. 3, 6

[11] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *TPAMI*, 40(4):834–848, 2017. 2, 4

[12] Qiang Chen, Jian Wang, Chuchu Han, Shan Zhang, Zexian Li, Xiaokang Chen, Jiahui Chen, Xiaodi Wang, Shuming Han, Gang Zhang, et al. Group detr v2: Strong object detector with encoder-decoder pretraining. *arXiv preprint arXiv:2211.03594*, 2022. 7

[13] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, pages 1597–1607. PMLR, 2020. 1, 2, 3, 4

[14] 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. 3, 6

[15] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *CVPR*, 2021. 3

[16] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *ECCV*, 2020. 3

[17] Zhe Chen, Yuchen Duan, Wenhai Wang, Junjun He, Tong Lu, Jifeng Dai, and Yu Qiao. Vision transformer adapter for dense predictions. *arXiv preprint arXiv:2205.08534*, 2022. 17

[18] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. *arXiv preprint arXiv:2112.01527*, 2021. 17

[19] MMSegmentation Contributors. MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. https://github.com/open-mmlab/mmsegmentation, 2020. 17

[20] Zihang Dai, Hanxiao Liu, Quoc V Le, and Mingxing Tan. Coatnet: Marrying convolution and attention for all data sizes. *NeurIPS*, 34:3965–3977, 2021. 2

[21] 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, 8, 16

[22] 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. 6, 16

[23] 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. 3, 5

[24] 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. In *ICLR*, 2020. 2, 6, 8, 14

[25] Robert M French. Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences*, 3(4):128–135, 1999. 6

[26] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, pages 580–587, 2014. 2, 4

[27] 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. In *NeurIPS*, 2020. 1, 2, 3, 4, 14

[28] 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. 4

[29] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *CVPR*, pages 5356–5364, 2019. 6, 8

[30] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022. 1, 2, 3, 4, 6, 7, 8, 14, 17
[31] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020. 3, 4, 14
[32] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 2
[33] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 3
[34] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006. 3
[35] 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. 14
[36] Ashish Jaiswal, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee, and Fillia Makedon. A survey on contrastive self-supervised learning. *Technologies*, 9(1):2, 2020. 2
[37] 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*, pages 4904–4916. PMLR, 2021. 1, 2, 3, 4
[38] Sebastian Kalkowski, Christian Schulze, Andreas Dengel, and Damian Borth. Real-time analysis and visualization of the yfcc100m dataset. In *Proceedings of the 2015 workshop on community-organized multimodal mining: opportunities for novel solutions*, pages 25–30, 2015. 6, 8
[39] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *NeurIPS*, 33:18661–18673, 2020. 3, 5
[40] Feng Li, Hao Zhang, Shilong Liu, Lei Zhang, Lionel M Ni, Heung-Yeung Shum, et al. Mask dino: Towards a unified transformer-based framework for object detection and segmentation. *arXiv preprint arXiv:2206.02777*, 2022. 7
[41] Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *AAAI*, 2020. 3
[42] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019. 3
[43] Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. *arXiv preprint arXiv:2203.16527*, 2022. 8, 17
[44] Feng Liang, Yangguang Li, and Diana Marculescu. Supmae: Supervised masked autoencoders are efficient vision learners. *arXiv preprint arXiv:2205.14540*, 2022. 3, 5
[45] Tingting Liang, Xiaojie Chu, Yudong Liu, Yongtao Wang, Zhi Tang, Wei Chu, Jingdong Chen, and Haibin Ling. Cbnet: A composite backbone network architecture for object detection. *IEEE TIP*, 2022. 16, 17
[46] 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*, pages 740–755. Springer, 2014. 2, 6, 8
[47] 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*, pages 12009–12019, 2022. 2, 6, 7
[48] 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. 2
[49] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *CVPR*, 2022. 2
[50] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*, 2019. 3
[51] Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens Van Der Maaten. Exploring the limits of weakly supervised pretraining. In *ECCV*, pages 181–196, 2018. 3
[52] Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. Slip: Self-supervision meets language-image pretraining. In *ECCV*, 2022. 3, 8
[53] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 3
[54] 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. 3, 14
[55] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PMLR, 2021. 1, 2, 3, 4, 7, 8, 14
[56] 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, 3, 6
[57] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *ICCV*, pages 8430–8439, 2019. 16, 17
[58] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, pages 2556–2565, 2018. 3
[59] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. In *CVPR*, pages 15638–15650, 2022. 3, 5
[60] Mannat Singh, Laura Gustafson, Aaron Adcock, Vinicius de Freitas Reis, Bugra Gedik, Raj Prateek Kosaraju, Dhruv

Mahajan, Ross Girshick, Piotr Dollár, and Laurens van der Maaten. Revisiting weakly supervised pre-training of visual perception models. In *CVPR*, pages 804–814, 2022. 3

[61] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. In *ICLR*, 2020. 3

[62] Chen Sun, Fabien Baradel, Kevin Murphy, and Cordelia Schmid. Learning video representations using contrastive bidirectional transformer. *arXiv preprint arXiv:1906.05743*, 2019. 3

[63] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *CVPR*, 2019. 3

[64] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *EMNLP*, 2019. 3

[65] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *ICML*, pages 6105–6114. PMLR, 2019. 1, 2

[66] Chenxin Tao, Honghui Wang, Xizhou Zhu, Jiahua Dong, Shiji Song, Gao Huang, and Jifeng Dai. Exploring the equivalence of siamese self-supervised learning via a unified gradient framework. In *CVPR*, 2022. 3

[67] Chenxin Tao, Xizhou Zhu, Weijie Su, Gao Huang, Bin Li, Jie Zhou, Yu Qiao, Xiaogang Wang, and Jifeng Dai. Siamese image modeling for self-supervised vision representation learning. In *CVPR*, 2023. 3, 4, 5, 8, 14, 17

[68] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016. 3, 6, 16

[69] Changyao Tian, Wenhai Wang, Xizhou Zhu, Jifeng Dai, and Yu Qiao. Vl-ltr: Learning class-wise visual-linguistic representation for long-tailed visual recognition. In *ECCV*, pages 73–91. Springer, 2022. 4

[70] Yuandong Tian, Xinlei Chen, and Surya Ganguli. Understanding self-supervised learning dynamics without contrastive pairs. In *ICML*, pages 10268–10278. PMLR, 2021. 3

[71] Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000. 4

[72] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *ICML*, pages 10347–10357. PMLR, 2021. 2

[73] Hugo Touvron, Matthieu Cord, and Herve Jegou. Deit iii: Revenge of the vit. *arXiv preprint arXiv:2204.07118*, 2022. 2, 8

[74] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 30, 2017. 6

[75] Andreas Veit, Maximilian Nickel, Serge Belongie, and Laurens van der Maaten. Separating self-expression and visual content in hashtag supervision. In *CVPR*, pages 5919–5927, 2018. 3

[76] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *ICML*, pages 1096–1103, 2008. 3

[77] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *ICML*, pages 23318–23340. PMLR, 2022. 3

[78] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442*, 2022. 2, 3, 6, 7, 17, 18

[79] Wenhai Wang, Jifeng Dai, Zhe Chen, Zhenhang Huang, Zhiqi Li, Xizhou Zhu, Xiaowei Hu, Tong Lu, Lewei Lu, Hongsheng Li, Xiaogang Wang, and Yu Qiao. Internimage: Exploring large-scale vision foundation models with deformable convolutions. In *CVPR*, 2023. 2, 6, 7, 16

[80] Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Dense contrastive learning for self-supervised visual pre-training. In *CVPR*, 2021. 3, 14

[81] Yixuan Wei, Han Hu, Zhenda Xie, Zheng Zhang, Yue Cao, Jianmin Bao, Dong Chen, and Baining Guo. Contrastive learning rivals masked image modeling in fine-tuning via feature distillation. *arXiv preprint arXiv:2205.14141*, 2022. 2, 3, 6, 7, 14

[82] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *CVPR*, 2018. 3

[83] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *ECCV*, pages 418–434, 2018. 2, 4, 17

[84] Enze Xie, Jian Ding, Wenhai Wang, Xiaohang Zhan, Hang Xu, Peize Sun, Zhenguo Li, and Ping Luo. Detco: Unsupervised contrastive learning for object detection. In *ICCV*, pages 8392–8401, 2021. 3

[85] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *CVPR*, pages 10687–10698, 2020. 2

[86] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *CVPR*, pages 1492–1500, 2017. 2

[87] 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*, pages 9653–9663, 2022. 3

[88] Chenglin Yang, Siyuan Qiao, Qihang Yu, Xiaoding Yuan, Yukun Zhu, Alan Yuille, Hartwig Adam, and Liang-Chieh Chen. Moat: Alternating mobile convolution and attention brings strong vision models. *arXiv preprint arXiv:2210.01820*, 2022. 7

[89] 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. 3, 5

[90] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *ICCV*, pages 6023–6032, 2019. 2, 5, 6

[91] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *ICML*, pages 12310–12320. PMLR, 2021. 1, 2, 3, 4, 14

[92] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *CVPR*, 2022. 2

[93] Chaoning Zhang, Chenshuang Zhang, Junha Song, John Seon Keun Yi, Kang Zhang, and In So Kweon. A survey on masked autoencoder for self-supervised learning in vision and beyond. *arXiv preprint arXiv:2208.00173*, 2022. 2

[94] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017. 2, 5, 6

[95] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022. 16, 17

[96] Haotian Zhang, Pengchuan Zhang, Xiaowei Hu, Yen-Chun Chen, Liunian Harold Li, Xiyang Dai, Lijuan Wang, Lu Yuan, Jenq-Neng Hwang, and Jianfeng Gao. Glipv2: Unifying localization and vision-language understanding. *arXiv preprint arXiv:2206.05836*, 2022. 3, 7

[97] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *CVPR*, pages 633–641, 2017. 6, 8

[98] 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*, 127(3):302–321, 2019. 2

[99] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. *International Conference on Learning Representations (ICLR)*, 2022. 8