Correlational Image Modeling for Self-Supervised Visual Pre-Training

Wei Li, Jiahao Xie, Chen Change Loy
S-Lab, Nanyang Technological University
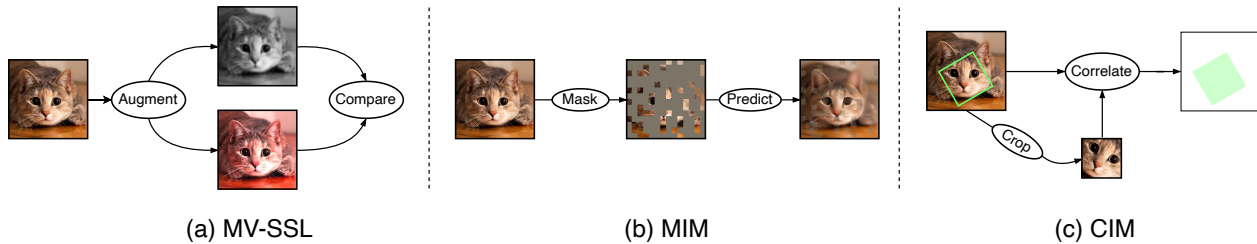{wei.l, jiahao003, ccloy}@ntu.edu.sg

Figure 1. **Schematic of pretext tasks** in self-supervised visual pre-training. (**a**) Multi-View Self-Supervised Learning (MV-SSL) follows an ***augment-and-compare*** paradigm. (**b**) Masked Image Modeling (MIM) conducts a ***mask-and-predict*** pretext task within a single view. (**c**) Correlational Image Modeling (CIM) formulates a novel ***crop-and-correlate*** scheme.

Abstract

*We introduce **C**orrelational **I**mage **M**odeling (CIM), a novel and surprisingly effective approach to self-supervised visual pre-training. Our CIM performs a simple pretext task: we randomly crop image regions (exemplars) from an input image (context) and predict correlation maps between the exemplars and the context. Three key designs enable correlational image modeling as a nontrivial and meaningful self-supervisory task. First, to generate useful exemplar-context pairs, we consider cropping image regions with various scales, shapes, rotations, and transformations. Second, we employ a bootstrap learning framework that involves online and target encoders. During pre-training, the former takes exemplars as inputs while the latter converts the context. Third, we model the output correlation maps via a simple cross-attention block, within which the context serves as queries and the exemplars offer values and keys. We show that CIM performs on par or better than the current state of the art on self-supervised and transfer benchmarks. Code is available at https://github.com/weivision/Correlational-Image-Modeling.git.*

1. Introduction

Recent advances in self-supervised visual pre-training have shown great capability in harvesting meaningful representations from hundreds of millions of—often eas-ily accessible—*unlabeled* images. Among existing pre-training paradigms, Multi-View Self-Supervised Learning (MV-SSL) [8–12, 21, 23] and Masked Image Modeling (MIM) [2, 22, 54, 68] are two leading methods in the self-supervised learning racetrack, thanks to their nontrivial and meaningful *self-supervisory pretext tasks*.

MV-SSL follows an ***augment-and-compare*** paradigm (Figure 1(a)) – randomly transforming an input image into two augmented views and then comparing two different views in the representation space. Such an instance-wise discriminative task is rooted in *view-invariant* learning [43], *i.e.*, changing views of data does not affect the conveyed information. On the contrary, following the success of Masked Language Modeling (MLM) [16], MIM conducts a ***mask-and-predict*** pretext task within a single view (Figure 1(b)) – removing a proportion of random image patches and then learning to predict the missing information. This simple patch-wise generative recipe enables Transformer-based deep architectures [17] to learn generalizable representations from unlabeled images.

Beyond ***augment-and-compare*** or ***mask-and-predict*** pretext tasks in MV-SSL and MIM, in this paper, we endeavor to investigate another simple yet effective paradigm for self-supervised visual representation learning. We take inspiration from visual tracking [70] in computer vision that defines the task of estimating the motion or trajectory of a target object (*exemplar*) in a sequence of scene images (*contexts*). To cope with challenging factors such as scale variations, deformations, and occlusions, one typical tracking pipeline

is formulated as maximizing the correlation between the specific *exemplar* and holistic *contexts* [3,5,46,52]. Such simple correlational modeling can learn meaningful representations in the capability of both localization and discrimination, thus making it appealing to serve as a promising pretext task for self-supervised learning.

Training a standard correlational tracking model, however, requires access to numerous labeled data, which is unavailable in unsupervised learning. Also, the task goal of visual tracking is intrinsically learning toward one-shot object detection—demanding rich prior knowledge of objectness—while less generic for representation learning. Therefore, it is nontrivial to retrofit supervised correlational modeling for visual tracking into a useful self-supervised pretext task.

Driven by this revelation, we present a novel ***crop-and-correlate*** paradigm for self-supervised visual representation learning, dubbed as **C**orrelational **I**mage **M**odeling (`CIM`). To enable correlational modeling for effectively self-supervised visual pre-training, we introduce three key designs. First, as shown in Figure 1(c), we randomly crop image regions (treated as *exemplars*) with various scales, shapes, rotations, and transformations from an input image (*context*). The corresponding correlation maps can be derived from the exact crop regions directly. This simple cropping recipe allows us to easily construct the *exemplar-context* pairs together with ground-truth correlation maps without human labeling cost. Second, we employ a bootstrap learning framework that is comprised of two networks: an online encoder and a target encoder, which, respectively, encode *exemplars* and *context* into latent space. This bootstrapping effect works in a way that the model learns to predict the spatial correlation between the updated representation of *exemplars* and the slow-moving averaged representation of *context*. Third, to realize correlational learning, we introduce a correlation decoder built with a cross-attention layer and a linear predictor, which computes queries from *context*, with keys and values from *exemplars*, to predict the corresponding correlation maps.

Our contributions are summarized as follows: **1)** We present a simple yet effective pretext task for self-supervised visual pre-training, characterized by a novel unsupervised correlational image modeling framework (`CIM`). **2)** We demonstrate the advantages of our `CIM` in learning transferable representations for both ViT and ResNet models that can perform on par or better than the current state-of-the-art MIM and MV-SSL learners while improving model robustness and training efficiency. We hope our work can motivate future research in exploring new useful pretext tasks for self-supervised visual pre-training.

2. Related Work

Unsupervised pretext tasks play the fundamental role in self-supervised representation learning. Beyond *augment-and-compare* and *mask-and-predict*, a series of different unsupervised pretext tasks have been studied in the literature. For instance, Noroozi *et al.* [37] train a context-free network without human annotation by solving Jigsaw puzzles, further developed in a very recent work [71] by predicting positions from content images. Bojanowski *et al.* [4] propose to learn discriminative features via predicting noise. Gidaris *et al.* [19] treat the 2D rotation of an image as a supervisory signal. Zhang *et al.* [72] follow this work to predict general affine transformations. All these initiatives are proven less effective than the state-of-the-art MIM and MV-SSL approaches in large-scale visual pre-training.

Multi-view self-supervised learning approaches [8–12, 21, 23, 36, 58, 65] are highly successful in learning representations over the past few years. These methods depend on an *augment-and-compare* pretext task that models similarity and dissimilarity between two or more augmented views in an embedding space. Thus, MV-SSL greatly relies on data augmentations and Siamese networks [6]. There have been several general strategies for comparing augmented views. Most contrastive approaches, such as SimCLR [9], MoCo [10, 12, 23] measure both positive and negative pairs via cosine distance. On the contrary, BYOL [21] and Sim-Siam [11] rely only on positive pairs. Beyond contrastive learning, SwAV [7] resorts to online clustering and predicts cluster assignments of different views. In addition, there is another line of research in MV-SSL that extends the main focus of *global* representations to *dense* representations [1, 24, 29, 38–40, 53, 55, 60, 61, 64, 67, 69].

Masked image modeling follows a *mask-and-predict* pretext task, which is inspired by the successful masked language modeling (MLM) approaches in the NLP community, such as BERT [16] and RoBERTa [33]. Two key steps can be identified in a typical MIM pipeline: i) *how to mask*, ii) *what to predict*. In terms of *how to mask*, most MIM approaches, such as BEiT [2], MAE [22] and SimMIM [68], extend the *mask-word* recipe in MLM to randomly mask image patches in the spatial domain. Recent works consider other corruptions to replace the normal patch-masking process. For example, Xie *et al.* [63] investigate corruption operations (downsample, blur, and noise) in low-level image processing tasks and present a unified *mask-frequency* recipe. Similarly, other degradation forms are studied in Tian *et al.* [44], including zoom, distortion, and decolorization. Besides, Fang *et al.* [18] employ an auxiliary generator to corrupt the input images. As to *what to predict*, beyond default raw pixels [22, 68], several other reconstruction targets are proposed, *e.g.*, hand-crafted or deep features [54], low or high frequencies [32, 63], and discrete tokens [2].

Correlational modeling is the crucial process in visual tracking [70], aiming to predict a dense set of matching confidence for a target object. The seminal work of Correlation Filter [5] and its end-to-end Siamese-based variants [3, 28, 46, 51, 52]

learn to distinguish targets from background images via convolution (*i.e.*, cross-correlation). Recently, Transformer-based trackers [13, 14, 34, 42, 62] employ a cross-attention mechanism to model the correlation between target objects and backgrounds. These promising correlation-based trackers motivate us to investigate the effectiveness of correlational modeling in the context of self-supervised visual pre-training. Notably, some unsupervised and self-supervised trackers [27, 41, 50, 57, 73] normally conduct training on synthetic datasets without labeling. While similarly considering unsupervised or self-supervised training, our work significantly differs from these unsupervised and self-supervised trackers. We will discuss the differences in Section 3.3.

3. Approach

Our correlational image modeling (`CIM`) is a simple yet effective self-supervised pre-training approach. As illustrated in Figure 2, we formulate a ***crop-and-correlate*** pretext task that crops a random image region (*exemplar*) from an input image (*context*) and predicts the correlation map between the *exemplar* and *context*. Our `CIM` consists of four components: cropping strategy, encoder, decoder, and loss function. In the following, we first introduce correlation operations in Section 3.1. We subsequently detail each component of `CIM` in Section 3.2. We finally discuss the relation of our `CIM` with the unsupervised visual tracking task in Section 3.3.

3.1. Preliminary: Correlation Operation

Given an *exemplar* image $z \in \mathbb{R}^{h_z \times w_z \times 3}$ along with a typically larger *context* image $c \in \mathbb{R}^{h_c \times w_c \times 3}$, a correlation operation between the *exemplar* and *context* images is defined as follows:

$$f(z, c) = f_\theta(z) \star f_\theta(c) + b\mathbb{1}, \qquad (1)$$

where $\star$ denotes a correlation operator and f_θ is a backbone model to extract corresponding representations. Here, $b\mathbb{1}$ represents a signal that takes value b in every location. Conceptually, it means that a dense similarity of two sets is measured in a 2D fashion. For instance, in standard Siamese-based trackers [3, 28, 46, 51], $\star$ is normally instantiated as a 2D convolution operator, in which the exemplar feature $f_\theta(z)$ takes the role of convolutional kernels, sliding over the spatial region of context feature $f_\theta(c)$. For Transformer-based trackers [13, 14, 34, 42, 62], a cross-attention layer combines information from two images to generate a merged representation, which selectively highlights the hotspots in the *context*.

3.2. Correlational Image Modeling

Cropping strategy. To enable effective correlational image modeling for self-supervised visual pre-training, we propose
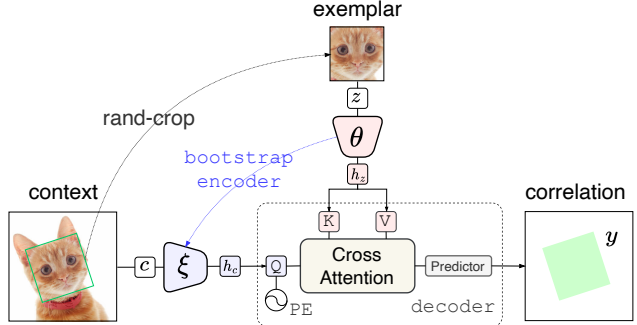

Figure 2. The overview of our proposed `CIM` pre-training framework. Given an image c (*context*), we crop a random region z (*exemplar*) within *context* c. The *context* and *exemplar* images are separately passed through a target encoder f_ξ and an online encoder f_θ to obtain latent representations h_c and h_z, which are further fed into a lightweight decoder with a cross-attention layer and a linear predictor to predict the correlation map y.

a random cropping strategy to construct *exemplar-context* image pairs. Specifically, as shown in Figure 3, given an original image $x \in \mathbb{R}^{H \times W \times 3}$, we obtain a *context* image $c \in \mathbb{R}^{m \times m \times 3}$ with a square shape by first randomly cropping a sub-region followed by a resizing operation. Then, we repeat the *crop-and-resize* process to generate a square *exemplar* image $z \in \mathbb{R}^{n \times n \times 3}$ from the *context* in consideration of three aspects: scale, shape, and rotation. To control the scale of an *exemplar*, we calculate the areas of both the cropping region and *context* and compute the scale ratio r_0. The shape of the cropping region is determined by the height and width ratio r_1. Also, we measure the rotation degree α between the cropping region and *context* image. By random sampling the values of r_0, r_1, and α, we can obtain *exemplars* with various scales, shapes, and rotations. The corresponding correlation map $y \in \{0, 1\}^{m \times m}$ can be derived from the cropping region easily. We further add different transformations to each *exemplar* image to increase the data diversity. We will study the effects of different cropping strategies in the experiment section.

Encoder. The goal of `CIM` is to learn useful representations with a backbone model f_θ in Equation 1, such that f_θ can be generalized to downstream tasks. Both ViT and CNN architectures can be applied as the encoder for `CIM`. For reliable pre-training, we employ a bootstrap encoder that consists of an online network θ and a target network ξ, which share the same backbone architecture. Given a pair of *exemplar* and *context* (z and c), we obtain the corresponding representations:

$$h_z = f_\theta(z); h_c = f_\xi(c) \qquad (2)$$

via the online network f_θ and target network f_ξ, respectively. The parameters ξ of the target network are updated from the online network θ with an exponential moving average
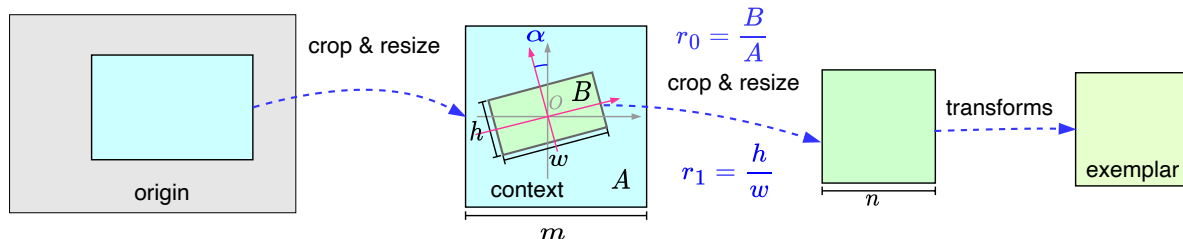
Figure 3. The procedure to generate an *exemplar-context* pair for CIM. We control the scale, shape, and rotation of an *exemplar* image by randomly sampling $r_0 = \frac{B}{A}, r_1 = \frac{h}{w}$ and, α, where A and B are the areas of cropping region ($h \times w$) and *context* image ($m \times m$).

policy [30]:

$$\xi = \tau\xi + (1 - \tau)\theta, \qquad (3)$$

where $\tau \in [0, 1]$ denotes a target decay rate. As a result, the online network f_θ is responsible for learning the representations to deal with various scales, shapes, and rotations for the *exemplar* images. For efficient training, we consider cropping multiple *exemplars* for each *context*, and all the cropped *exemplars* can be grouped together into one forward-backward processing.

Decoder. To model the correlation between *exemplar* and *context* images, we design a lightweight cross-attention decoder, which is a general form of multi-head attention layer in Transformers [48]. To be specific, we first project the representations h_z and h_c to obtain the query, key, and value by linear mappings: $q = f_c(h_c) + \text{PE}; k = f_k(h_z); v = f_v(h_z)$. A positional encoding (PE) is added to the query for better decoding. The reason why we use the *context* as a query rather than the *exemplar* is that the output correlation map is of the shape determined by the *context* input, not the *exemplar*. Then we can calculate the weighted representation for the *exemplar* and *context* pair as follows:

$$a = \text{CrossAttention}(h_c, h_z) = \text{Softmax}\left(\frac{qk^\top}{\sqrt{d}}\right)v. \qquad (4)$$

After that, we compute the correlational representation with *layernorm* and *multilayer perceptron* modules:

$$h = h_c + a + \text{MLP}(\text{LN}(h_c + a)). \qquad (5)$$

Finally, the output correlation map $\widehat{y}$ is obtained via a linear predictor and an *upsampling* operation[1]:

$$\widehat{y} = \text{Upsample}(f_p(h)). \qquad (6)$$

Loss function. In practice, to optimize the overall CIM model, we simply minimize a binary cross-entropy loss:

$$\mathcal{L}(\widehat{y}, y) = -\frac{1}{m \times m} \sum_{i=1}^{m \times m} y_i \log(\widehat{y}_i) + (1 - y_i) \log(1 - \widehat{y}_i), \qquad (7)$$

between predicted correlation map $\widehat{y}$ and ground-truth y.

---

[1]https://pytorch.org/docs/stable/generated/torch.nn.Upsample.html

3.3. Relation to Unsupervised Visual Tracking

Our CIM is generally related to studies on unsupervised and self-supervised visual tracking [27,41,50,57,73]. These works explore effective training cues to bypass the need for extensive annotated data for training deep tracking models. Typically, temporal consistency or correspondence in videos is leveraged as a cue. Several modeling techniques have been proposed, including forward and backward consistency [50], progressive training [57], multi-task learning [41], and memory augmenting [27]. Despite the different strategies, all these unsupervised and self-supervised trackers mainly focus on learning task-specific representations for visual tracking from unlabeled videos. Thus, it is infeasible to apply these trackers with temporal modeling on still images, in which such temporal information does not exist. On the contrary, the goal of our CIM is to learn generic representations from unlabeled data with the transferable ability to downstream tasks. Therefore, we formulate correlational modeling in a more general form and develop it as a useful pretext task for self-supervised visual pre-training.

4. Experiments

4.1. Main Properties

To understand the unique properties of CIM, we conduct ablation studies on ImageNet-200 [47], a smaller subset of the ImageNet-1K dataset [15]. For all ablation experiments, we choose ViT-Base (ViT-B/16) as the default backbone and follow a common setting used in existing works [2, 22]: *300-epoch self-supervised pre-training without labels and 100-epoch supervised end-to-end fine-tuning*, to evaluate the quality of learned representations. For a fair comparison, we tailor the resolutions of *context* and *exemplar* as 160×160 and 64×64, respectively. By default, we crop six *exemplars* for each input image (*context*), in order to match with the standard 224×224 input size.[2] More detailed pre-training and fine-tuning recipes are described in the supplementary material. We present our observations as follows:

---

[2]For ViT-Base (ViT-B/16) with 16×16 patch size, our configuration of one *context* (160×160) with six *exemplars* (64×64) contains 196 image patches in total, which is equivalent to an image with the size of 224×224.

Table 1. **Ablations of cropping strategy** for `CIM` with ViT-B/16 on ImageNet-200.

(a) **Crop scale**.

Scale	Ratio r_0	Top-1 acc (%)
scratch	-	77.79
MoCo v3 [12]	-	89.60
MAE [22]	-	89.03
fixed	$r_0 = 0.16$	87.57
random	$r_0 < 0.16$	87.25
random	$r_0 > 0.16$	89.39
random	$r_0 \in (0, 1.0)$	**89.48**

(b) **Crop shape**.

Shape	Ratio r_1	Top-1 acc (%)
square	1.0	89.48
rectangle	$[3/4, 4/3]$	89.55
rectangle	$[1/2, 2/1]$	89.59
rectangle	$[1/3, 3/1]$	**89.70**
rectangle	$[1/4, 4/1]$	89.66

(c) **Rotation**.

Rotation α	Top-1 acc (%)
$0°$	89.70
$[-45°, 45°]$	**89.97**
$[-90°, 90°]$	89.91
$[-135°, 135°]$	89.19
$[-180°, 180°]$	89.07

(d) **Transformation**.

context	*exemplar*	Top-1 acc (%)
✗	✗	89.97
✓	✗	90.01
✗	✓	**90.12**
✓	✓	90.12

Table 2. **Ablations of encoder designs** for `CIM` with ViT-B/16 on ImageNet-200.

Bootstrap	Update	Top-1 acc (%)
scratch	-	77.79
✗	shared	89.91
✓	$\xi \to \theta$	89.05
✓	$\theta \to \xi$	**90.12**

Cropping strategy. We investigate how different cropping strategies will affect our `CIM` in self-supervised representation learning. We consider four aspects of cropping *exemplars*, *i.e.*, scale, shape, rotation, and transformation:

(i) `Scale`: As shown in Table 1a, we study the scale factor of *exemplars* while keeping the shape and rotation fixed, *i.e.*, square shape and $0°$ rotation. We first consider cropping with fixed scale ratio $r_0 = \frac{64 \times 64}{160 \times 160} = 0.16$ and then study three random scale ratio schemes: small scale ($r_0 < 0.16$), large scale ($r_0 > 0.16$), and both small and large scales $r_0 \in (0, 1.0)$. All these entries perform significantly better than the baseline, *i.e.*, training from scratch without pretraining. The random cropping policy that covers both small and large scales $r_0 \in (0, 1.0)$ performs best. This indicates that adding variation on the scale ratio of cropping *exemplar* images can help our `CIM` to learn better representations.

(ii) `Shape`: In Table 1b we further study the crop shape of *exemplars*. Given that deep architectures (CNNs and ViTs) are more easily to process rectangle inputs, we extend the square in previous studies to the rectangle with height/width ratio r_1. Other non-rectangle shapes (*e.g.*, triangles and circles) are beyond our study. We find that expanding the sampling range of shape ratio r_1 as $[1/3, 3/1]$ can boost the performance upon the square entry. However, if a larger range of $[1/4, 4/1]$ is applied, no further performance gain can be obtained. All these experiments suggest that the shape of *exemplars* is also a useful factor in our cropping strategy.

(iii) `Rotation`: Table 1c shows the influence of the rotation degree α between the cropping *exemplars* and *context* image. We conduct rotation experiments upon the previous best entry in Table 1b. The optimal sampling range of α is $[-45°, 45°]$, which means adding a relatively smaller degree of rotation is helpful for our `CIM`, while large rotation degree such as $[-180°, 180°]$ would bring a negative effect. Our consideration of rotation is strategically different from previous predicting image rotations [19] in that we treat rotation as a type of augmentation rather than a supervisory signal. Therefore, a reasonably small degree of rotation can improve our `CIM` pre-training.

(iv) `Transformation`: Table 1d studies the influence of data transformations on our `CIM` pre-training. We consider random data transformations including horizontal flipping, Gaussian blurring, color jittering, grayscale, and solarization. We can observe that only adding transformations on *exemplars* while keeping *context* images unaltered works best for our `CIM`. This can be explained by our bootstrap encoder design: the online network encode *exemplars* while the offline network processes the *context*, as a result, *exemplars* are more responsible for affecting model training. Note that our transformation for *exemplar-context* pairs is different from existing MV-SSL methods such as MoCo v3 [12] and DINO [8], in which two transformed views are conceptually identical, thus can be swapped during training.

Encoder design. Our `CIM` encoder follows a bootstrap design with *exemplar* and *context* images encoded by the online network f_θ and target network f_ξ, respectively. As studied in Table 2, we first notice that learning with a shared encoder can achieve significantly better performance than the training-from-scratch baseline. We further evaluate two bootstrapping designs: 1) $\xi \to \theta$ training update from *context* to *exemplar*, and 2) $\theta \to \xi$ training update from *exemplar* to *context*. Obviously, we can find that $\theta \to \xi$ entry performs better than both the shared and $\xi \to \theta$ entries. This validates
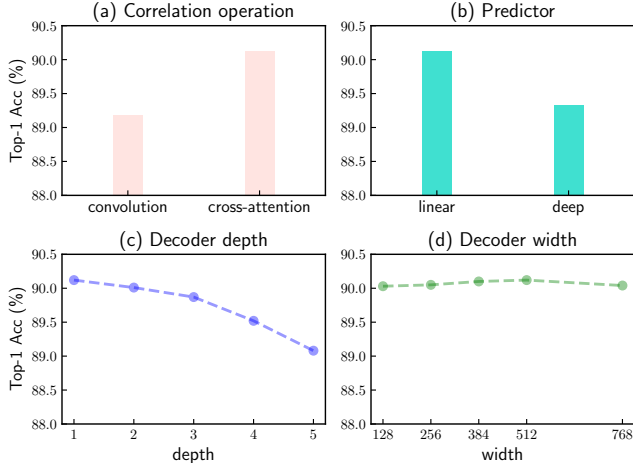
Figure 4. **Ablations of decoder designs** for `CIM` with ViT-B/16 on ImageNet-200.

the rationality of our bootstrap encoder design, which is also consistent with our finding of transformation in Table 1d: our `CIM` benefits more from learning with *exemplars* than *context* images.

Decoder design. Our `CIM` decoder plays a key role in correlational modeling. We study our decoder designs as follows:

(i) `Correlation operation`: Figure 4 (a) compares two correlation operations commonly used in existing deep tracking models. As we introduced in Section 3.1, when applying convolution as the correlation operation, *exemplars* features are served as the kernels and convolve with *context* features, in which local correlations are computed in each kernel window. Differently, as formulated in Equation 4, a cross-attention layer models global correlation between *exemplar* and *context* images. The cross-attention entry can yield up to 1% improvement over the convolution entry. This suggests that modeling global correlation is better for `CIM`.

(ii) `Predictor`: In Figure 4 (b), we study the network design of the final predictor. A simple linear layer followed by an `Unsample` operation works well for our `CIM`. While a deep predictor with three deconvolution layers cannot bring further gain but degrade `CIM` training. This suggests a lightweight predictor may force `CIM` to better representations in the encoder, while a heavy predictor is more specialized for predicting accurate correlation maps in the decoder but less relevant for representation learning.

(iii) `Depth`: Figure 4 (c) varies the decoder depth (number of cross-attention layers). Interestingly, a *single* cross-attention layer works best for our `CIM` training. Adding more layers brings no training gain for correlation modeling, similar to our observation in previous predictor designs.

(iv) `Width`: Figure 4 (d) studies the decoder width (number of heads in each cross-attention layer). We set 512-d by default, which performs well for our `CIM`. Increasing or decreasing the layer width does not cause significant ac-

Table 3. **Ablations of loss fuctions** for `CIM` with ViT-B/16 on ImageNet-200.

Loss Function	CE	BCE	MSE	Focal
Top-1 acc (%)	**90.12**	89.28	87.68	77.35

Table 4. **Comparisons with visual tracking works** with ViT-B/16 on ImageNet-200.

Method	Scratch	SiamFC	SiamRPN	TransTrack	CIM
Top-1 (%)	77.79	89.09	89.02	89.54	**90.12**

curacy improvement or degradation. The decoder depth is less influential for improving representation learning for our `CIM`.

Overall, our `CIM` decoder is lightweight. It has only one cross-attention layer with a width of 512-d and a linear layer for final prediction. As such, our `CIM` is efficient in model pre-training.

Loss function. We study the influence of different loss functions for our `CIM` optimization in Table 3. Given that our `CIM` predicts binary correlation maps, we compare typical loss functions for dense predictions, including cross-entropy (CE), balanced cross-entropy (BCE) [66], mean squared error (MSE), and Focal loss [31]. A standard cross-entropy performs best for our correlation modeling. This property is dramatically different from deep visual tracking models [3, 46, 52] and related unsupervised trackers [27, 41, 50, 57, 73], which clearly benefit from proper dense objectives. This can be explained by the difference in task goals between `CIM` and visual tracking: our `CIM` focuses on learning transferable representations by correlation modeling, whereas deep visual trackers demand task-specific representations in favor of better dense predictions.

4.2. Comparisons with Visual Tracking Models

Our `CIM` is inspired by the correlation modeling in supervised visual tracking models. Our proposed cropping strategy can generate useful *exemplar-context* pairs that are also suitable for training supervised visual tracking models. We train three representative trackers using ViT-B/16 as the backbone: SiamFC [3], SiamRPN [28], and TransTrack [13], with generated *exemplar-context* pairs on ImageNet-200. Following the same pre-training and fine-tuning setting in previous ablation studies, we evaluate the quality of learned representations, as summarized in Table 4. We can observe that: (1) Owing to the *exemplar-context* pairs generated by our cropping strategy, all three trackers can learn good representations that perform better than the scratch baseline. (2) Based on SiamFC, SiamRPN introduces an additional detection head for bounding box prediction, which brings no performance gain. (3) TransTrack works better than both SiamFC and SiamRPN. This is due in large part to the benefi-

Table 5. **ImageNet-1K top-1 fine-tuning accuracy** of self-supervised models using ViT-S/16 and ViT-B/16 as the encoder. All entries are on an image size of 224×224. We use the actual processed images/views to measure the effective pre-training epochs [75]. Scratch indicates the supervised baseline in [45]. † denotes results are reproduced using the official code.

Method	Pre-train Data	Pretext Task	Tokenizer	Epochs	ViT-S	ViT-B
Scratch [45]	-	-	-	-	79.9	81.8
MP3 [71]	IN-1K	Jigsaw	-	100	-	81.9
MoCo v3 [12]	IN-1K	MV-SSL	-	1200	81.4	83.2
DINO [8]	IN-1K	MV-SSL	-	1600	81.5	82.8
BEiT [2]	IN-1K+DALL-E	MIM	dVAE	300	81.3	82.9
SimMIM [68]†	IN-1K	MIM	-	300	80.9	82.9
MAE [22]†	IN-1K	MIM	-	300	80.6	82.9
CIM	IN-1K	CIM	-	300	81.6	83.1

cial global correlation modeling provided by cross-attention layers, in comparison with local correlations computed by convolution operations in SiamFC and SiamRPN. (4) Our CIM clearly surpasses these visual tracking works, showing the advantages of our encoder and decoder designs for effective correlational modeling.

4.3. Comparisons with Previous SSL Methods

Our CIM is a general framework that can learn meaningful representations for both ViT and CNN architectures, unlike state-of-the-art methods such as MAE [22].

ViT. In Table 5 we first compare the fine-tuning results of ViT-S/16 and ViT-B/16 models with self-supervised pretraining on ImageNet-1k. Following previous works [2, 22], we fine-tune ViT-S/16 for 200 epochs, and ViT-B/16 for 100 epochs. More detailed pre-training and fine-tuning configurations are described in the supplementary material. Compared with previous MV-SSL works [8, 12], such as MoCo v3 [12], our CIM can achieve highly comparable performances (83.1 vs. 83.2), while enjoying significantly fewer epochs of pretraining (300 vs. 1200). Compared with previous MIM works [2, 22, 68], using the same 300 epochs of pre-training, our CIM can achieve better performances with both ViT-S/16 and ViT-B/16 models.

ResNet-50. We further demonstrate that our CIM can effectively pre-train the classic ResNet architecture. During pre-training, we simply apply the same ViT pre-training configurations for ResNet-50. To evaluate the pre-trained representations, we generally follow the state-of-the-art vanilla ResNet "training-from-scratch" recipe in RSB [56]. We present detailed fine-tuning settings in the supplementary material. The evaluation results compared to the state-of-the-art methods are summarized in Table 6. Due to the architectural difference between ViT and CNN models, we observe performance degeneration of some MIM and MV-SSL pretraining methods, such as SimMIM [68], MoCo v2 [10], and SimSiam [11]. Compared with the best MV-SSL method,

Table 6. **ImageNet-1K top-1 fine-tuning accuracy** of self-supervised models using ResNet-50 as the encoder. † denotes results are reproduced using the official code.

Method	Pretext Task	Epochs	Top-1 acc (%)
Fine-tuning for 100 epochs			
RSB A3 [56]	-	-	78.1
SimMIM [68]†	MIM	300	77.7
CIM	CIM	300	78.6
Fine-tuning for 300 epochs			
RSB A2 [56]	-	-	79.8
SimSiam [11]	MV-SSL	400	79.1
MoCo v2 [10]	MV-SSL	400	79.6
SimCLR [9]	MV-SSL	800	79.9
BYOL [21]	MV-SSL	400	80.0
SwAV [7]	MV-SSL	600	80.1
SimMIM [68]†	MIM	300	79.5
CIM	CIM	300	80.1

SwAV [7], our CIM is faster (300 vs. 600).

Overall, our CIM is a simple yet effective approach that can perform on par or better than existing MV-SSL and MIM methods with both ViT and ResNet models.

4.4. Transfer Learning on Semantic Segmentation

To evaluate the transferability of the pre-trained representations by our CIM, we further conduct end-to-end finetuning on the ADE20K [74] semantic segmentation benchmark. Following the same setup in BEiT [2], we fine-tune the pre-trained ViT-B/16 model as the backbone in UperNet [59] for 160K iterations, with an input resolution of 512×512. As summarized in Table 7, our CIM can achieve highly competitive performance compared with other representative self-supervised learners. This demonstrates the effectiveness of our proposed *crop-and-correlate* pretext task in learning transferable representations.

Figure 5. **Visualization** of *exemplar-context* images in company with both ground-truth and predicted correlation maps for `CIM`.

Table 7. **ADE20K semantic segmentation** of ViT-B/16 models.

Method	Pre-train Data	Pretext Task	mIoU (%)
Supervised [45]	IN-1K w/ labels	-	45.3
MoCo v3 [12]	IN-1K	MV-SSL	47.2
DINO [8]	IN-1K	MV-SSL	46.8
BEiT [2]	IN-1K+DALL-E	MIM	47.7
MAE [22]	IN-1K	MIM	48.1
`CIM`	IN-1K	CIM	48.1

4.5. Robustness Evaluation

We further evaluate the robustness of our models on six benchmarks that cover the challenges of adversarial attacks, common corruption, and out-of-distribution. For adversarial attack, we evaluate the adversarial examples on ImageNet-A [26], along with generated examples by FGSM [20] and PGD [35] attackers on ImageNet-1K validation set. For data corruption, we test corrupted images on ImageNet-C [25]. In terms of out-of-distribution input, we consider images with distribution shifts from ImageNet-R [25] and ImageNet-Sketch [49]. Specifically, we directly evaluate the models fine-tuned on original ImageNet-1K (ViT-B/16 in Table 5 and ResNet-50 in Table 6) without further fine-tuning on each robustness validation set. The results are summarized in Table 8. For both ViT and ResNet architectures, our `CIM` consistently outperforms the state-of-the-art self-supervised learners for model robustness.

4.6. Visualization

In Figure 5, we visualize the *exemplar-context* images generated by our proposed cropping strategy on the ImageNet-1K validation set. The predicted correlation maps are obtained via the ViT-B/16 model pre-trained on the ImageNet-1K train set using our `CIM` in Section 4.3. No further pre-training or fine-tuning is conducted. We can observe that these predicted correlation maps match closely with the corresponding ground-truth correlations, under various scales, shapes, rotations, and transformations. The results

Table 8. **Robustness evaluation on six robustness benchmarks.** We report top-1 accuracy except for IN-C which uses the mean corruption error (mCE). The original ImageNet top-1 fine-tuning results are also appended for reference. The best results are in **bold**, and the second best results are underlined.

Method	Robustness Benchmarks						Orig.
	FGSM	PGD	IN-C (↓)	IN-A	IN-R	IN-SK	
ViT-B/16 model results							
Scratch [56]	46.3	21.2	**48.5**	28.1	44.7	32.0	81.8
MAE [22]	38.9	11.2	52.3	31.5	48.3	33.8	82.9
`CIM`	**47.4**	**22.7**	49.3	**30.3**	**48.6**	**35.3**	**83.1**
ResNet-50 model results							
Scratch [56]	**20.2**	**3.4**	77.0	6.6	36.0	25.0	78.1
SimMIM [68]	16.8	2.1	77.0	5.7	34.9	24.2	77.7
`CIM`	19.4	2.5	**73.5**	**8.5**	**37.4**	27.2	**78.6**

demonstrate the effectiveness of self-supervised correlation modeling in our `CIM` for unseen data. More visualized examples are provided in the supplementary material.

5. Conclusion

In this work, we present `CIM`, a novel pretext task for self-supervised visual pre-training. Unlike existing MV-SSL and MIM approaches, `CIM` considers correlation modeling in visual tracking as a useful pre-training paradigm. We build a generic self-supervised correlational modeling framework by proposing three unique designs, including a cropping strategy, bootstrap encoder, and correlation decoder. Extensive experiments on transfer learning and robustness evaluation with visual recognition tasks show that our `CIM` can efficiently and effectively learn meaningful representations from unlabeled images.

References

[1] Yutong Bai, Xinlei Chen, Alexander Kirillov, Alan Yuille, and Alexander C Berg. Point-level region contrast for object detection pre-training. In *CVPR*, 2022. 2

[2] Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021. 1, 2, 4, 7, 8

[3] Luca Bertinetto, Jack Valmadre, Joao F Henriques, Andrea Vedaldi, and Philip HS Torr. Fully-convolutional siamese networks for object tracking. In *ECCV*, pages 850–865. Springer, 2016. 2, 3, 6

[4] Piotr Bojanowski and Armand Joulin. Unsupervised learning by predicting noise. In *ICML*, pages 517–526. PMLR, 2017. 2

[5] David S Bolme, J Ross Beveridge, Bruce A Draper, and Yui Man Lui. Visual object tracking using adaptive correlation filters. In *CVPR*, pages 2544–2550. IEEE, 2010. 2

[6] Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. Signature verification using a" siamese" time delay neural network. *NeurIPS*, 6, 1993. 2

[7] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *NeurIPS*, 2020. 2, 7

[8] 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. 1, 2, 5, 7, 8

[9] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020. 1, 2, 7

[10] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 1, 2, 7

[11] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *CVPR*, 2021. 1, 2, 7

[12] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *ICCV*, 2021. 1, 2, 5, 7, 8

[13] Xin Chen, Bin Yan, Jiawen Zhu, Dong Wang, Xiaoyun Yang, and Huchuan Lu. Transformer tracking. In *CVPR*, pages 8126–8135, 2021. 3, 6

[14] Yutao Cui, Cheng Jiang, Limin Wang, and Gangshan Wu. Mixformer: End-to-end tracking with iterative mixed attention. In *CVPR*, pages 13608–13618, 2022. 3

[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*, pages 248–255. IEEE, 2009. 4

[16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019. 1, 2

[17] 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. 1

[18] Yuxin Fang, Li Dong, Hangbo Bao, Xinggang Wang, and Furu Wei. Corrupted image modeling for self-supervised visual pre-training. *arXiv preprint arXiv:2202.03382*, 2022. 2

[19] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *ICLR*, 2018. 2, 5

[20] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 8

[21] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. In *NeurIPS*, 2020. 1, 2, 7

[22] 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, 7, 8

[23] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020. 1, 2

[24] Olivier J Hénaff, Skanda Koppula, Jean-Baptiste Alayrac, Aaron van den Oord, Oriol Vinyals, and João Carreira. Efficient visual pretraining with contrastive detection. In *ICCV*, 2021. 2

[25] 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. 8

[26] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *CVPR*, 2021. 8

[27] Zihang Lai, Erika Lu, and Weidi Xie. Mast: A memory-augmented self-supervised tracker. In *CVPR*, pages 6479–6488, 2020. 3, 4, 6

[28] Bo Li, Junjie Yan, Wei Wu, Zheng Zhu, and Xiaolin Hu. High performance visual tracking with siamese region proposal network. In *CVPR*, pages 8971–8980, 2018. 2, 3, 6

[29] Zhaowen Li, Yousong Zhu, Fan Yang, Wei Li, Chaoyang Zhao, Yingying Chen, Zhiyang Chen, Jiahao Xie, Liwei Wu, Rui Zhao, et al. Univip: A unified framework for self-supervised visual pre-training. In *CVPR*, 2022. 2

[30] Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015. 4

[31] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, pages 2980–2988, 2017. 6

[32] Hao Liu, Xinghua Jiang, Xin Li, Antai Guo, Deqiang Jiang, and Bo Ren. The devil is in the frequency: Geminated gestalt autoencoder for self-supervised visual pre-training. *arXiv preprint arXiv:2204.08227*, 2022. 2

[33] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized

bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 2

[34] Fan Ma, Mike Zheng Shou, Linchao Zhu, Haoqi Fan, Yilei Xu, Yi Yang, and Zhicheng Yan. Unified transformer tracker for object tracking. In *CVPR*, pages 8781–8790, 2022. 3

[35] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017. 8

[36] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *CVPR*, 2020. 2

[37] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *ECCV*, 2016. 2

[38] Pedro O O Pinheiro, Amjad Almahairi, Ryan Benmalek, Florian Golemo, and Aaron C Courville. Unsupervised learning of dense visual representations. *NeurIPS*, 33:4489–4500, 2020. 2

[39] Byungseok Roh, Wuhyun Shin, Ildoo Kim, and Sungwoong Kim. Spatially consistent representation learning. In *CVPR*, 2021. 2

[40] Ramprasaath R Selvaraju, Karan Desai, Justin Johnson, and Nikhil Naik. Casting your model: Learning to localize improves self-supervised representations. In *CVPR*, 2021. 2

[41] Qiuhong Shen, Lei Qiao, Jinyang Guo, Peixia Li, Xin Li, Bo Li, Weitao Feng, Weihao Gan, Wei Wu, and Wanli Ouyang. Unsupervised learning of accurate siamese tracking. In *CVPR*, pages 8101–8110, 2022. 3, 4, 6

[42] Zikai Song, Junqing Yu, Yi-Ping Phoebe Chen, and Wei Yang. Transformer tracking with cyclic shifting window attention. In *CVPR*, pages 8791–8800, 2022. 3

[43] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning. In *NeurIPS*, 2020. 1

[44] Yunjie Tian, Lingxi Xie, Jiemin Fang, Mengnan Shi, Junran Peng, Xiaopeng Zhang, Jianbin Jiao, Qi Tian, and Qixiang Ye. Beyond masking: Demystifying token-based pre-training for vision transformers. *arXiv preprint arXiv:2203.14313*, 2022. 2

[45] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *ICML*, 2021. 7, 8

[46] Jack Valmadre, Luca Bertinetto, Joao Henriques, Andrea Vedaldi, and Philip HS Torr. End-to-end representation learning for correlation filter based tracking. In *CVPR*, pages 2805–2813, 2017. 2, 3, 6

[47] Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, Marc Proesmans, and Luc Van Gool. Scan: Learning to classify images without labels. In *ECCV*, pages 268–285. Springer, 2020. 4

[48] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 4

[49] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. In *NeurIPS*, 2019. 8

[50] Ning Wang, Yibing Song, Chao Ma, Wengang Zhou, Wei Liu, and Houqiang Li. Unsupervised deep tracking. In *CVPR*, pages 1308–1317, 2019. 3, 4, 6

[51] Qiang Wang, Li Zhang, Luca Bertinetto, Weiming Hu, and Philip HS Torr. Fast online object tracking and segmentation: A unifying approach. In *CVPR*, pages 1328–1338, 2019. 2, 3

[52] Qiang Wang, Yun Zheng, Pan Pan, and Yinghui Xu. Multiple object tracking with correlation learning. In *CVPR*, pages 3876–3886, 2021. 2, 6

[53] Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Dense contrastive learning for self-supervised visual pre-training. In *CVPR*, 2021. 2

[54] Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. *arXiv preprint arXiv:2112.09133*, 2021. 1, 2

[55] Fangyun Wei, Yue Gao, Zhirong Wu, Han Hu, and Stephen Lin. Aligning pretraining for detection via object-level contrastive learning. 2021. 2

[56] Ross Wightman, Hugo Touvron, and Hervé Jégou. Resnet strikes back: An improved training procedure in timm. *arXiv preprint arXiv:2110.00476*, 2021. 7, 8

[57] Qiangqiang Wu, Jia Wan, and Antoni B Chan. Progressive unsupervised learning for visual object tracking. In *CVPR*, pages 2993–3002, 2021. 3, 4, 6

[58] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *CVPR*, 2018. 2

[59] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *ECCV*, 2018. 7

[60] Tete Xiao, Colorado J Reed, Xiaolong Wang, Kurt Keutzer, and Trevor Darrell. Region similarity representation learning. In *ICCV*, 2021. 2

[61] 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*, 2021. 2

[62] Fei Xie, Chunyu Wang, Guangting Wang, Yue Cao, Wankou Yang, and Wenjun Zeng. Correlation-aware deep tracking. In *CVPR*, pages 8751–8760, 2022. 3

[63] Jiahao Xie, Wei Li, Xiaohang Zhan, Ziwei Liu, Yew Soon Ong, and Chen Change Loy. Masked frequency modeling for self-supervised visual pre-training. In *ICLR*, 2023. 2

[64] Jiahao Xie, Xiaohang Zhan, Ziwei Liu, Yew Soon Ong, and Chen Change Loy. Unsupervised object-level representation learning from scene images. In *NeurIPS*, 2021. 2

[65] Jiahao Xie, Xiaohang Zhan, Ziwei Liu, Yew-Soon Ong, and Chen Change Loy. Delving into inter-image invariance for unsupervised visual representations. *IJCV*, 2022. 2

[66] Saining Xie and Zhuowen Tu. Holistically-nested edge detection. In *ICCV*, pages 1395–1403, 2015. 6

[67] Zhenda Xie, Yutong Lin, Zheng Zhang, Yue Cao, Stephen Lin, and Han Hu. Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning. In *CVPR*, 2021. 2

[68] 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. *arXiv preprint arXiv:2111.09886*, 2021. 1, 2, 7, 8

[69] Ceyuan Yang, Zhirong Wu, Bolei Zhou, and Stephen Lin. Instance localization for self-supervised detection pretraining. In *CVPR*, 2021. 2

[70] Alper Yilmaz, Omar Javed, and Mubarak Shah. Object tracking: A survey. *ACM Computing Surveys (CSUR)*, 38(4):13–es, 2006. 1, 2

[71] Shuangfei Zhai, Navdeep Jaitly, Jason Ramapuram, Dan Busbridge, Tatiana Likhomanenko, Joseph Yitan Cheng, Walter Talbott, Chen Huang, Hanlin Goh, and Joshua Susskind. Position prediction as an effective pretraining strategy. In *ICML*, 2022. 2, 7

[72] Liheng Zhang, Guo-Jun Qi, Liqiang Wang, and Jiebo Luo. Aet vs. aed: Unsupervised representation learning by auto-encoding transformations rather than data. In *CVPR*, 2019. 2

[73] Jilai Zheng, Chao Ma, Houwen Peng, and Xiaokang Yang. Learning to track objects from unlabeled videos. In *ICCV*, pages 13546–13555, 2021. 3, 4, 6

[74] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *IJCV*, 127(3):302–321, 2019. 7

[75] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. Image bert pre-training with online tokenizer. In *ICLR*, 2022. 7