UniVIP: A Unified Framework for Self-Supervised Visual Pre-training

Zhaowen Li [1,2]   Yousong Zhu [1]   Fan Yang[3]   Wei Li[3]   Chaoyang Zhao [1,4]   Yingying Chen [1]
Zhiyang Chen [1,2]   Jiahao Xie [5]   Liwei Wu [3]   Rui Zhao [3,7]   Ming Tang [1]   Jinqiao Wang [1,2,6]

National Laboratory of Pattern Recognition, Institute of Automation, CAS, Beijing, China[1]
School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China[2]
SenseTime Research[3]   Development Research Institute of Guangzhou Smart City, Guangzhou, China[4]
S-Lab, Nanyang Technological University[5]   Peng Cheng Laboratory, Shenzhen, China[6]
Qing Yuan Research Institute, Shanghai Jiao Tong University, Shanghai, China[7]

{zhaowen.li,yousong.zhu,chaoyang.zhao,yingying.chen,zhiyang.chen,tangm,jqwang}@nlpr.ia.ac.cn
{yangfan1,liwei1,wuliwei,zhaorui}@sensetime.com   jiahao003@ntu.edu.sg

Abstract

Self-supervised learning (SSL) holds promise in leveraging large amounts of unlabeled data. However, the success of popular SSL methods has limited on single-centric-object images like those in ImageNet and ignores the correlation among the scene and instances, as well as the semantic difference of instances in the scene. To address the above problems, we propose a Unified Self-supervised Visual Pre-training (UniVIP), a novel self-supervised framework to learn versatile visual representations on either single-centric-object or non-iconic dataset. The framework takes into account the representation learning at three levels: 1) the similarity of scene-scene, 2) the correlation of scene-instance, 3) the discrimination of instance-instance. During the learning, we adopt the optimal transport algorithm to automatically measure the discrimination of instances. Massive experiments show that UniVIP pre-trained on non-iconic COCO achieves state-of-the-art transfer performance on a variety of downstream tasks, such as image classification, semi-supervised learning, object detection and segmentation. Furthermore, our method can also exploit single-centric-object dataset such as ImageNet and outperforms BYOL by 2.5% with the same pre-training epochs in linear probing, and surpass current self-supervised object detection methods on COCO dataset, demonstrating its universality and potential.

1. Introduction

Deep learning has shown excellent performance on various computer vision tasks [12, 20, 21, 25, 32, 47] using labels. Self-supervised learning (SSL) of visual representa-
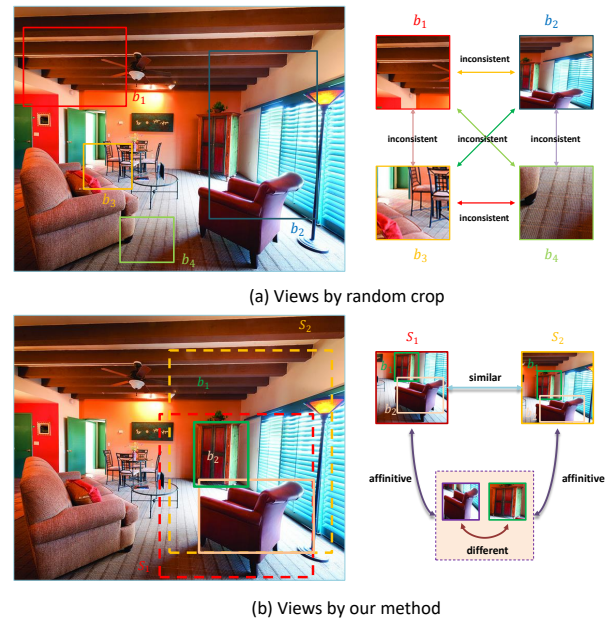

(a) Views by random crop

(b) Views by our method

Figure 1. **Visualization of different cropped views on COCO dataset.** (a) Images from COCO contain multiple instances, thus different random crops might represent different semantic meanings and are not satisfied with the semantic consistency assumption. (b) Our method for unified self-supervised learning. The two scene views are created with overlapping regions while the overlapping regions contain multiple instances.

tion aims at capturing salient feature representation without relying on human annotations. Recently, contrastive learning [1, 4–7, 17, 18, 23, 41] based SSL has proved impressive results on a number of downstream tasks, largely narrowing the gap between unsupervised and supervised learning and even surpassing the supervised counterpart. These state-of-

the-art (SOTA) methods build upon the pretext task called instance discrimination, which regards different views of a single image as the same instance and its objective is simply to learn a feature representation that discriminates among images. Hence, the basic assumption of these methods is that the pre-training data should have the property of semantic consistency [3, 27], *i.e.*, the assumption highly relies on the single-centric-object data such as those in ImageNet [12]. Nevertheless, it is infeasible for complex natural images since they usually consist of multiple instances as shown in Fig 1(a). Some work [17,29,34] naively extend the off-the-shelf SSL methods from ImageNet to other datasets, like MS COCO [26], Places365 [52], and YFCC100M [36], yet they do not acquire satisfactory results.

It is known that multiple instances in the single natural image possess the co-occurrence relationship, and usually have different semantic meanings. Therefore, the models should have the ability to distinguish the semantics of different instances. However, it is still challenging to discriminate different instances residing in the single natural image when no instance annotations are available. Several region-level based methods [27, 30, 33] propose to leverage multiple local regions to pre-train models using non-iconic dataset, and achieve the success of the specific downstream task. Nevertheless, these region-level based methods do not explicitly distinguish different instances in the scene. In addition, their results of linear evaluation are inferior to the baseline, *i.e.*, these methods can not obtain versatile visual representations. Moreover, natural images have prior that the scene and instances in the scene have the semantic affinity since these instances correlate with the scene. Current SSL methods are not aware of the prior and do not encode the semantic affinity. Because of the above problems, the application scenario of these methods is limited. It is essential to design an effective learning paradigm to obtain versatile visual representations.

In this paper, we introduce a unified self-supervised pre-training framework, named UniVIP, to learn the visual representations by pre-training on either single-centric-object or non-iconic dataset. Specifically, we first exploit the unsupervised instance proposal method Selective Search [38] to generate candidate instances. Then, for each image, we create two scene views with overlapping regions containing instances to guarantee the global similarity, *i.e.*, the similarity of scene-scene, as much as possible, which will effectively alleviate the semantic inconsistency of different scene views. Moreover, to tackle the correlation of scene-instance, the generated instances are grouped to approximate the semantics of the corresponding scene views, guiding the network to learn a variety of instances in the image. In our UniVIP, the discrimination of instance-instance is formulated as the optimal matching problem among all candidate instances in overlapping regions and uses the op-

timal transport algorithm [15] to discriminate different instances in the scene. Our objective consists of the above three items, and different views of some scenes and instances obtained by our UniVIP are shown in Fig 1(b). It is noted that our framework is specially designed to learn versatile representations from natural images, and is able to fully leverage the prior of semantic affinity among the natural scene and instances in the scene, and explicitly distinguish co-occurrence instances.

Massive experiments in single-centric-object and non-iconic datasets prove that UniVIP can learn the versatile representations. In particular, our method outperforms the state-of-the-art by 2.3% top-1 classification accuracy with pre-training on COCO dataset for the ImageNet [12] linear evaluation protocol. Our 300-epoch UniVIP achieves 42.2 bbox mAP and 38.2 mask mAP using Mask R-CNN [20] on COCO detection and segmentation with $1\times$ schedule when pre-trained on ImageNet, and even surpasses the popular self-supervised object detection methods.

Overall, we make the following contributions:

- We proposed a Unified Self-supervised Representations Learning framework to effectively overcome the semantic inconsistency of random views in non-iconic images, and it can be pre-trained with any images.
- We proposed to simultaneously leverage the similarity of scene-scene, the correlation of scene-instance, and the discrimination of instance-instance to promote the performance of models effectively.
- Extensive experiments demonstrate the effectiveness and stronger generalization ability of our method. Specifically, the models pre-trained with UniVIP on single-centric-object and non-iconic datasets all outperform previous SOTA methods in multiple downstream tasks, such as image classification, semi-supervised learning, object detection and segmentation.

2. Related work

2.1. SSL on single-centric-object dataset

Currently, the most competitive pretext task for self-supervised visual representation learning is instance discrimination [1, 5–7, 17, 18]. The learning objective is simply to learn representations by distinguishing each image from others, and this approach has proved excellent performance on extensive downstream tasks, such as image classification [12, 21], object detection [25, 32] and segmentation [20]. MoCo [18] improves the training of instance discrimination methods by storing representations from a momentum encoder instead of the trained network. SimCLR [5] shows that the memory bank can be entirely replaced with the elements from the same batch if the batch is large enough. Meanwhile, BYOL [17] directly bootstraps the representations by attracting the different features from
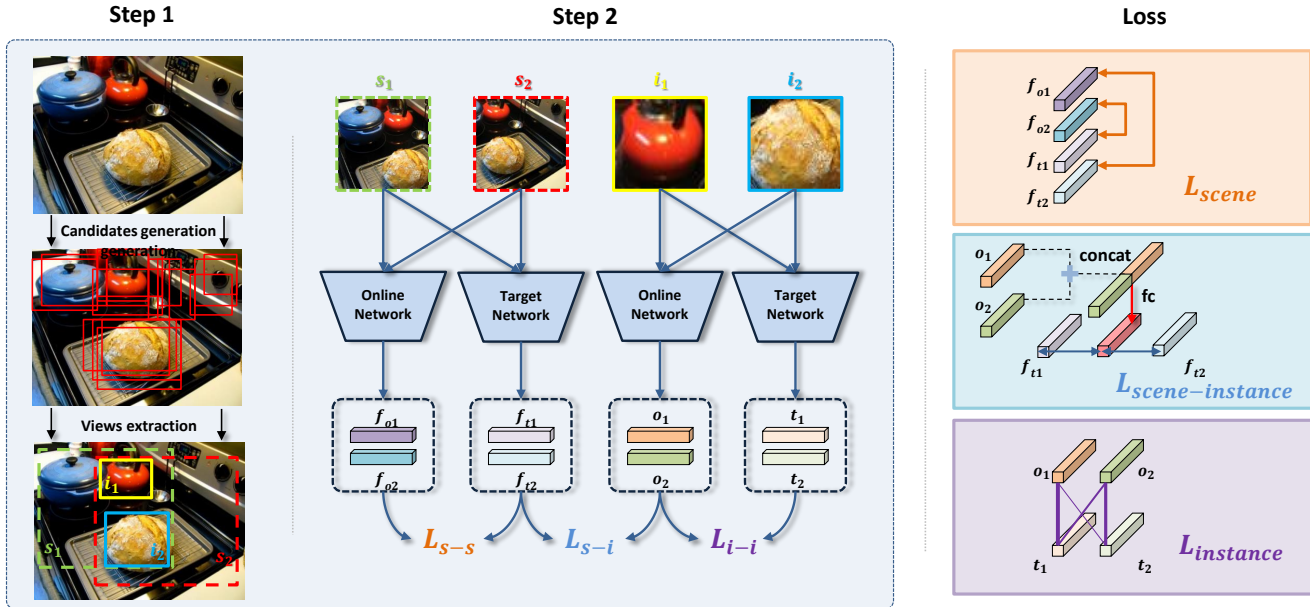
Figure 2. **The pipeline of UniVIP.** The non-iconic image is first extracted candidate instances using unsupervised object proposal algorithms selective search. Then, we create two views with overlapping regions and multiple instances in the overlapping regions from the image. The existence of overlapping regions can guarantee the scene's similarity. Here, we adopt two instances as an example. Furthermore, we feed the two scene views and two instances to the online and target networks, and obtain feature representations. Finally, we compute the similarity of scene-scene, the correlation of scene-instance, and the discrimination of instance-instance. The total loss function consists of the $\mathcal{L}_{scene}$, $\mathcal{L}_{scene\text{-}instance}$, and $\mathcal{L}_{instance}$.

the same instance. SwAV [1] maps the image features to a set of trainable prototype vectors and proposes a multi-crop strategy for self-supervised data augmentation. Moreover, Some works [2, 9] extend the pretext task to vision Transformer [10, 13] and achieve superior performance in image classification. Furthermore, MST [23] propose an attention-guided mask strategy to avoid masking crucial regions of images for self-supervised Transformer learning. Nevertheless, much of their progress has far been limited to single-centric-object pre-training data such as ImageNet [12], and may be infeasible when extended to non-iconic datasets.

2.2. SSL on natural scene

Purushwalkam *et al.* [31] point out that the advance of current SSL methods moderately comes from their usage of dataset bias of ImageNet. Also, they find that training MoCo on the less-biased COCO dataset does not get encouraging results. Moreover, HED [49] report degraded performance when training MoCo on COCO and PASCAL [14] datasets. Recently, MaskCo [51] also notices that the semantic consistency assumption of current SSL methods and proposes a contrastive mask prediction task for visual representation learning. Some preliminary work [17,29,34] naively extend the off-the-shelf constrastive learning methods from ImageNet to other datasets, like MS COCO [26], Places365 [52], and YFCC100M [36], yet they do not acquire satisfactory results since these datasets are not satisfied with the semantic consistency assumption, even though

the size of some datasets is orders of magnitude larger than ImageNet. Furthermore, DenseCL [30], Self-EMD [27], MaskCo and SCLR [33] leverage local regions of non-iconic images to pre-train the models, yet these methods only work on specific downstream task and can not acquire versatile visual representations. DnC [37] alternates between contrastive learning and clustering-based hard negative mining to train YFCC100M [36] and JFT-300M [35]. ORL [44] shows the impressive performance when pre-trained on MS COCO, but its three-stage method consumes much time, thus cannot support large-scale pre-training. These methods, however, are not aware of the semantic affinity between the scene and instances, and also ignore the semantic discrimination of different instances.

3. Approach

The pipeline of our proposed UniVIP is shown in Fig 2. We propose a unified visual self-supervised approach to learn versatile visual representations, which creatively integrates the scene similarity, the scene-instance semantic affinity, and the semantic discrimination of different instances. Here, we first review the basic instance discrimination method in 3.1. Then, the mechanism and effect of the scene similarity are explained in 3.2. Furthermore, we study the correlation of scene-instance in 3.3. Finally, the optimal transport algorithm is imported to promote the semantic discrimination of different instances, and the training function of our method is described in 3.4.

3.1. Preliminary

Without loss of generality, we adopt BYOL [17] as our basic self-supervised learning method, which can achieve state-of-the-art transfer performance. For each image x, BYOL first generates two views $x_1 \sim \mathcal{T}_1(x)$ and $x_2 \sim \mathcal{T}_1(x)$ under random data augmentation, which are then fed into the *online* network $f_\theta(x)$ and the *target* network $g_\xi(x)$ separately, parameterized by θ and ξ. Both online and target networks possess a neural network backbone and a projection head [6], which share the same architecture with different parameters. While the online network has the target predictor [17]. The parameters ξ of fixed network $f_\xi(x)$ is updated by the exponential moving average of the parameters θ of online network according to Eq (1), where decay rate $m \in [0,1]$ is the momentum and will increase to 1.0 until the end of training.

$$\xi = m * \xi + (1 - m) * \theta \tag{1}$$

Finally, BYOL maximizes the cosine similarity between the prediction of online network and the projected feature of target network as the scene-level consistency. The loss function is defined as Eq (2).

$$\mathcal{L}(x_1, x_2) \triangleq -\frac{\langle f_\theta(x_1), g_\xi(x_2) \rangle}{||f_\theta(x_1)||_2 \cdot ||g_\xi(x_2)||_2}, \tag{2}$$

3.2. Similarity of scene-scene

The semantic consistency assumption is almost always satisfied in the single-centric-object ImageNet, which is the highly curated pre-training dataset. However, the implicit assumption can not be scalable to the natural dataset with non-iconic images. The main reason why inconsistency happens in non-iconic images is that the two random views may be far away from each other.

Meanwhile, the instance annotations of the natural images are unavailable. Therefore, to acquire the candidate instances as the prior, we leverage the unsupervised instance proposal algorithms selective search to generate proposals for each image. To filter the certain number of redundancy of generated proposals, we set some pre-defined thresholds including the minimal scale, the range of aspect ratio, and the maximal intersection-over-union (IoU) among these instance-based regions. Considering the instances that exist in the natural scene, we create two scene views $\mathbf{s_1}$, $\mathbf{s_2}$ with the overlapping regions containing K identical instances. For ensuring that each image has K regions, we generate boxes by the *naive* strategy if the number of candidate instances in the overlapping regions is less than K. The naive strategy includes setting the minimum scale to 64 pixels, the range of aspect ratio is between $1/3$ and $3/1$, and the maximum IoU threshold is 0.5.

By constructing two scene views with overlapping regions, we transfer the semantic inconsistency of random

views to similarity of scene-scene in natural images. In particular, we feed the two scene views to the online and target network separately, and acquire the representations $\mathbf{f_{o1}}$, $\mathbf{f_{o2}}$, $\mathbf{f_{t1}}$, and $\mathbf{f_{t2}}$ to compute the symmetric loss as Eq (3), following BYOL [17].

$$\mathcal{L}_{\text{scene}} = \mathcal{L}(\mathbf{s_1}, \mathbf{s_2}) + \mathcal{L}(\mathbf{s_2}, \mathbf{s_1}) \tag{3}$$

3.3. Correlation of scene-instance

Concretely, natural images have prior that the scene and instances residing in it possess the semantic affinity since these instances correlate with the scene. Apparently, it is reasonable to argue that exploring the prior among the scene and the instances is conducive to learning more general feature representations. However, current un-/self-supervised learning methods do not take into account the existence of the correlation of scene-instance. To research the correlation in the un-/self-supervised learning field, the primary issue is how to measure the semantic affinity among the scene and instances. Due to the simplicity and effectiveness of the cosine similarity, UniVIP attempts to establish the semantic affinity among the scene and multiple instances by the cosine similarity. Specifically, we crop and resize each instance i_k in overlapping regions to 96×96, where $k = \{1, ..., K\}$, then we feed K instances into the online network and obtain K representations vectors $[\mathbf{o_1}, \mathbf{o_2}, ..., \mathbf{o_K}]$. Furthermore, we concatenate these representations, and linearly map the concatenated representation to the dimension of the scene's representation and obtain the final representation $\mathbf{I}$ as Eq (4).

$$\mathbf{I} = f_{\text{linear}}(\text{concat}(\mathbf{o_1}, \mathbf{o_2}, ..., \mathbf{o_K})) \tag{4}$$

Finally, we minimize the cosine distance of the scene view $\mathbf{s}$ and instances combination in the feature space as Eq (5) and argue that the measurement can explore the prior of scene-instance semantic affinity.

$$\mathcal{L}_{\text{affinity}}(\mathbf{s}, \mathbf{I}) \triangleq -\frac{\langle \mathbf{I}, g_\xi(\mathbf{s}) \rangle}{||\mathbf{I}||_2 \cdot ||g_\xi(\mathbf{s})||_2} \tag{5}$$

We also compute the symmetric views since the overlapping area of two views possesses the same instances. Thus, the semantic affinity can be explored accroding to Eq (6).

$$\mathcal{L}_{\text{scene-instance}} = \mathcal{L}_{\text{affinity}}(\mathbf{s_1}, \mathbf{I}) + \mathcal{L}_{\text{affinity}}(\mathbf{s_2}, \mathbf{I}) \tag{6}$$

3.4. Discrimination of instance-instance

In subsection 3.3, we increase the affinity among the scene and instances, yet can not ensure that extracted features in each instance can be distinguished from other instances. Moreover, the contrastive loss [18] demands many negative samples, while the number of instances in the non-iconic image is limited, which can not meet the demand.

Therefore, in this subsection, we formulate the discrimination of instance-instance as an optimal transport problem. Here, we first describe the concept of optimal transport, then introduce how to apply the optimal transport to train the models for learning visual feature representations. Finally, we establish the training function of UniVIP.

Optimal transport. The form of Optimal Transport (OT) can be described as the following problem: supposing that a set of M suppliers are required to transport goods to a set of N demanders. The m-th supplier holds b_m units of goods while the n-th demander needs a_n units of goods. The cost per unit transported from supplier m to demander n is denoted by c_{mn}. The goal of optimal transport algorithm is to find a transportation plan $\tilde{\mathcal{Y}} = \{y_{m,n}|m = 1, 2, ...M, n = 1, 2, ...N\}$, according to which all goods from suppliers can be transported to demanders at a minimal transportation cost as Eq (7).

$$\min_y \quad \sum_{m=1}^{M}\sum_{n=1}^{N} c_{mn}y_{mn}.$$
$$\text{s.t.} \quad \sum_{m=1}^{M} y_{mn} = a_n, \quad \sum_{n=1}^{N} y_{mn} = b_m, \qquad (7)$$
$$\sum_{m=1}^{M} b_m = \sum_{n=1}^{N} a_n,$$
$$y_{mn} \geq 0, \quad m = 1, 2, ...M, n = 1, 2, ...N.$$

OT for semantic discrimination. For feeding the candidate instances in overlapping regions to the online network, there have a set of feature vectors $[\mathbf{o_1}, \mathbf{o_2}, ..., \mathbf{o_K}]$, and each vector o_m can be seen as a node in the set. Moreover, we also feed these instances to the target network. Similarly, the set of feature vectors $[\mathbf{t_1}, \mathbf{t_2}, ..., \mathbf{t_K}]$ can be acquired by the target network. Following the original optimal transport formulation in Eq (7), the cost per unit transported from supplier feature node $\mathbf{o_m}$ to demander node $\mathbf{t_n}$ is defined as Eq (8). Thus, the nodes with similar representations tend to generate fewer transport cost between each other while the nodes with irrelevant representations tend to generate more transport cost. The similarity of each pair of instances can be represented as the optimal matching cost between two sets of vectors.

$$c_{mn} = 1 - \frac{\mathbf{o}_m^T \mathbf{t}_n}{\|\mathbf{o}_m\|\|\mathbf{t}_n\|} \qquad (8)$$

Moreover, the marginal weights a_m and b_n are defined as Eq (9) and Eq (10), where the function $max(\cdot)$ ensures the weights are always non-negative.

$$b_m = \max\{\mathbf{o}_m^T \cdot \frac{\mathbf{f_{t1}} + \mathbf{f_{t2}}}{2}, 0\} \qquad (9)$$

$$a_n = \max\{\mathbf{t}_n^T \cdot \frac{\mathbf{f_{o1}} + \mathbf{f_{o2}}}{2}, 0\} \qquad (10)$$

Method	Arch	Batch size	pre-training # epochs	ImageNet Top-1 Acc
Random [16]	R-50	-	-	13.7
Pre-training on ImageNet:				
Supervised [29]	R-50	256	90	75.9
Pre-training on MS COCO:				
SimCLR [5]		512	800	50.9
MoCo v2		512	800	55.1
BYOL [17]	R-50	512	800	57.8
ORL [44]		512	800	59.0
UniVIP (ours)		512	800	**60.2**
Pre-training on COCO+:				
BYOL [17]		512	800	59.6
ORL [44]	R-50	512	800	60.7
UniVIP (ours)		512	800	**63.0**
Pre-training on ImageNet:				
InstDis [41]		256	200	58.5
MoCo [18]		256	200	60.6
CPC v2 [22]		512	200	63.8
SimCLR [5]		4096	200	66.5
MoCo v2 [7]		256	200	67.7
SwAV [1]	R-50	4096	200	69.1
SimSiam [8]		256	200	70.0
InfoMin [50]		256	200	70.1
BYOL [17]		4096	200	70.6
UniVIP (ours)		4096	200	**73.1**
UniVIP (ours)		4096	300	**74.2**

Table 1. **Comparison of SOTA self-supervised learning methods.** UniVIP is pre-trained on non-iconic COCO(+), and single-centric-object ImageNet. The COCO(+) and ImageNet pre-trained UniVIP all achieve state-of-the-art performance than previous SSL methods. For evaluation, a linear classifier is trained on ImageNet, we report top-1 accuracy on the ImageNet `val` set.

Following [15], we address the Eq (7) by a fast iterative solution, named Sinkhorn-Knopp [11], and acquire the optimal matching flows $\tilde{\mathcal{Y}}$. Then we can compute the discrimination of instance-instance as Eq (11). Here, the loss would be minimized only if the representations of each instance are similar to itself and dissimilar to other instances, *i.e.*, the instance can be distinguished from other instances.

$$\mathcal{L}_{\text{instance}}(\mathbf{O}, \mathbf{T}) \triangleq -\sum_{m=1}^{K}\sum_{n=1}^{K} \frac{\mathbf{o}_m^T \mathbf{t}_n}{\|\mathbf{o}_m\|\|\mathbf{t}_n\|} \tilde{y}_{mn} \qquad (11)$$

Finally, the total loss of our UniVIP is formulated as Eq (12), and each loss coefficient is equally weighted.

$$\mathcal{L}_{\mathbf{UniVIP}} = \mathcal{L}_{\text{scene}} + \mathcal{L}_{\text{scene-instance}} + \mathcal{L}_{\text{instance}} \qquad (12)$$

4. Experiments

Datasets. We first pre-train models on the COCO `train2017` set that contains ∼118k images. COCO contains more natural and diverse scenes in the wild.

Method	Pre-train data	AP^{bb}	AP^{mk}
Rand Init [18]	-	31.0	28.5
Supervised [18]	ImageNet	38.9	35.4
SimCLR [5]	COCO	37.0(-1.9)	33.7(-1.7)
MoCov2 [7]	COCO	38.5(-0.4)	34.8(-0.6)
BYOL [17]	COCO	39.5(+0.6)	35.6(+0.2)
ORL [44]	COCO	40.3(+1.4)	36.3(+1.9)
Ours	COCO	**40.8**(+1.9)	**36.8**(+1.4)
BYOL [17]	COCO+	40.0(+1.1)	36.2(+0.8)
ORL [44]	COCO+	40.6(+1.7)	36.7(+1.3)
Ours	COCO+	**41.1**(+2.2)	**37.1**(+1.7)
InsDis [41]	ImageNet	37.4(-1.5)	34.1(-1.3)
PIRL [29]	ImageNet	37.5(-1.4)	34.0(-1.4)
SwAV [1]	ImageNet	38.5(-0.4)	35.4(0.0)
MoCo [18]	ImageNet	38.5(-0.4)	35.1(-0.3)
MoCov2 [7]	ImageNet	38.9(0.0)	35.5(+0.1)
BYOL [17]	ImageNet	40.7(+1.8)	36.9(+1.5)
Ours	ImageNet	**41.6**(+2.7)	**37.6**(+2.2)

Table 2. **Results of object detection and instance segmentation fine-tuned on COCO with** $1\times$ **schedule**. We adopt Mask R-CNN R50-FPN, and report the bounding box AP and mask AP on COCO `val2017`. The COCO(+) pre-trained methods are trained for 800 epochs while the ImageNet pre-trained methods are trained for 200 epochs. UniVIP outperforms all supervised and supervised conterparts. Green means increase and gray means decrease.

Then, we perform self-supervised learning on a larger "COCO+" dataset (COCO `train2017` set plus COCO `unlabeled2017` set) to verify whether our method can benefit from more unlabeled natural data. Finally, for verifying the unification of our method, we pre-train models on the single-centric-object ImageNet dataset. ImageNet `train` consists of ∼1.28 million training images.

Network architecture. Following recent un-/self-supervised methods [18,27,33,44,45,51], we adopt ResNet-50 [21] as the default backbone. The two branches are slightly different, with one using a regular backbone network, a regular projection head, and a prediction head, and the other using the momentum network with a moving average of the parameters of the regular backbone network and the projection head.

Image augmentations. In pre-training, the scene-level data augmentation strategy follows [17]. The setting of the scene image augmentation include: random horizontal flipping, random color jittering, random grayscale conversion, random Gaussian blurring, and solarization. For the instance-level augmentation, we directly crop the regions on the input images and resize them to 96×96. The subsequent augmentations exactly follow the scene-level ones.

Optimization. For pre-training on the COCO(+), we fully follow the setting of [44]. Specifically, we use the SGD optimizer with a weight decay of 0.0001 and a momentum of 0.9. We adopt the cosine learning rate decay schedule [28] with a base learning rate of 0.2, and the batch size is set to 512 by default. We train our models for 800 epochs with a warm-up period of 4 epochs. The exponential moving aver-

Method	Pre-train data	AP^{bb}	AP^{mk}
Rand Init [18]	-	36.7	33.7
Supervised [18]	ImageNet	40.6	36.8
BYOL [17]	COCO	40.8(+0.2)	37.0(+0.2)
Ours	COCO	**42.2**(+1.6)	**38.2**(+1.4)
BYOL [17]	COCO+	41.4(+0.8)	37.4(+0.6)
Ours	COCO+	**42.8**(+2.2)	**38.6**(+1.8)
MoCo [18]	ImageNet	40.8(+0.2)	36.9(+0.1)
MoCov2 [7]	ImageNet	40.9(+0.3)	37.0(+0.2)
BYOL [17]	ImageNet	42.2(+1.6)	38.0(+1.2)
Ours	ImageNet	**43.1**(+2.5)	**38.8**(+2.0)

Table 3. **Results of object detection and instance segmentation fine-tuned on COCO with** $2\times$ **schedule**. The COCO(+) pre-trained methods are trained for 800 epochs while the ImageNet pre-trained methods are trained for 200 epochs. UniVIP achieves the state-of-the-art performance.

Method	Epoch	Pre-train data	AP^{bb}	AP^{mk}
Rand Init	-	-	31.0	28.5
Supervised	90	ImageNet	38.9	35.4
Self-EMD [27]	800	COCO	39.3(+0.4)	-
DenseCL [39]	800	COCO	39.6(+0.7)	35.7(+0.3)
Resim-FPN [42]	200	ImageNet	39.8(+0.9)	36.0(+0.6)
DetCo [43]	200	ImageNet	40.1(+1.2)	36.4(+1.0)
DenseCL [39]	200	ImageNet	40.3(+1.4)	36.4(+1.0)
Resim-FPN [42]	400	ImageNet	40.3(+1.4)	36.4(+1.0)
BYOL [17]	300	ImageNet	40.9(+2.0)	37.0(+1.6)
SCRL [33]	200	ImageNet	41.0(+2.1)	37.5(+2.1)
SCRL [33]	1000	ImageNet	41.4(+2.5)	37.9(+2.5)
InsLoc-FPN [46]	200	ImageNet	41.4(+2.5)	37.1(+1.7)
InsLoc-FPN [46]	400	ImageNet	42.0(+3.1)	37.6(+2.2)
UniVIP	800	COCO	40.8(+1.9)	36.8(+1.4)
	800	COCO+	41.1(+2.2)	37.1(+1.7)
	200	ImageNet	41.6(+2.8)	37.6(+2.1)
	300	ImageNet	**42.2**(+3.3)	**38.2**(+2.8)

Table 4. **Comparison with other self-supervised object detection methods.** We report the results with $1\times$ schedule on COCO. UniVIP-200ep is on par with previous state-of-the-art, and UniVIP-300ep achieves the best performance.

age parameter m starts from 0.99 and is increased to 1 during training. Similarly, following the setting of BYOL [17], we pre-train the models on ImageNet.

4.1. Linear evaluation on ImageNet

We compare our method with other prevailing algorithms pre-trained on COCO, COCO+, and ImageNet datasets in Table 1. All these methods share the same backbone for fair comparison and evaluation by linear probing. For COCO dataset, our model achieves 60.2% top-1 accuracy with linear probing. It outperforms the previous best algorithm ORL by 1.2% at the same training epochs, and even approaches the performance of ORL with a much larger dataset (60.7% pre-trained on COCO+). Our algorithm relieves the need for a larger training dataset for self-supervised learning, and is able to obtain a decent result (60.2%) with only ∼118k images. Notably, our method achieves 63.0% top-1 accuracy and outperforms ORL by

2.3% when applied to COCO+, which means UniVIP becomes better using a larger dataset. It should be emphasized that UniVIP is a unified framework, which can also be applied with the single-centric-object dataset. Here we use the popular ImageNet as an example. With the same pre-training epochs, UniVIP outperforms BYOL by 2.5%, which is a state-of-the-art self-supervised learning method designed delicately for the single-centric-object dataset. These results indicate that UniVIP is a unified method, which can be pre-trained with any images for visual self-supervised representation learning.

4.2. Object detection and segmentation

COCO with $1\times$ and $2\times$ schedule. We perform object detection and segmentation experiments using Mask R-CNN detector [20] with R50-FPN [24] implemented in Detectron2 [40]. We fine-tune all layers end-to-end on COCO `train2017` set and evaluate on `val2017` ($\sim$5k images). The schedule is the default $1\times$ or $2\times$ following the same setup in [18, 50]. In Table 2, we show the results of the learned representation by different self-supervised methods pre-training on different datasets. For fair comparison, all these methods are pre-trained with the same epochs. It can be observed that our method achieves the best results with 40.8% bbox mAP (AP^{bb}) and 36.8% mask mAP (AP^{mk}) when pre-trained on COCO. It outperforms the ImageNet supervised counterpart by 1.9% and 1.4%, and BYOL results by 1.3% and 1.2%. Similarly, the COCO+ pre-trained UniVIP reaches 41.1% AP^{bb} and 37.1% AP^{mk}, which surpasses the supervised by 2.2% and 1.7%, and BYOL results by 1.1% and 0.9%. Notably, the results also show excellent performance when pre-trained on single-centric-object ImageNet. Our method yields 3.3% AP^{bb} and 2.8% AP^{mk} improvements over the supervised. Simultaneously, as shown in Table 3, the impressive performance is still retained when the pre-trained models are trained with longer epochs in downstream dense prediction tasks.

Comparison with current self-supervised object detection methods. It should be emphasized that our method also outperforms current state-of-the-art self-supervised object detection methods [27,33,39,42,43,46] as shown in Table 4, even though part methods [42, 46] pre-training with FPN [24]. Meanwhile, UniVIP acquires more gains with longer pre-training epochs, which shows the great potential of our framework. These phenomena indicate that learning the versatile visual representation can effectively enhance the transfer ability of models. Although our method requires about 35% computation than the BYOL, but the performance of 300-epoch UniVIP outperforms the 1000-epoch SCRL, which is an object detection self-supervised method based on BYOL. The results show that the computation of UniVIP brings a large gain.

COCO with longer training iterations. Moreover, in

pretrain	LR schedule		
	$1\times$	$2\times$	$6\times$
Random [50]	31.0	36.7	42.7
Supervised [50]	38.9(+7.9)	40.6(+3.9)	42.6(-0.1)
BYOL [17]	40.9(+9.9)	42.3(+5.6)	42.6(-0.1)
Ours	**42.2**(+11.2)	**43.5**(+6.8)	**44.0**(+1.3)

Table 5. **Results of object detection fine-tuned on COCO with longer training iterations.** The results of UniVIP-300ep object detection AP^{bb} on COCO `val2017` with training schedules from $1\times$ (90k iterations) to $6\times$ (540k iterations).

Method	Pre-train data	1% labels		10% labels	
		Top-1	Top-5	Top-1	Top-5
Random [44]	-	1.6	5.0	21.8	44.2
Supervised [48]	ImageNet	25.4	48.4	56.4	80.4
SimCLR [5]	COCO	23.4	46.4	52.2	77.4
MoCo v2 [7]	COCO	28.2	54.7	57.1	81.7
BYOL [17]	COCO	28.4	55.9	58.4	82.7
ORL [44]	COCO	31.0	58.9	60.5	84.2
Ours	COCO	**31.6**	**59.7**	**61.3**	**85.6**
BYOL [17]	COCO+	28.3	56.0	59.4	83.6
ORL [44]	COCO+	31.8	60.1	60.9	84.4
Ours	COCO+	**32.3**	**60.9**	**61.7**	**85.7**
BYOL [17]	ImageNet	51.5	76.7	64.7	86.6
Ours	ImageNet	**53.0**	**78.8**	**67.1**	**88.5**

Table 6. **Semi-supervised learning on ImageNet.** The COCO(+) pre-trained methods are trained for 800 epochs while the ImageNet pre-trained methods are trained for 200 epochs. We fine-tune all models with 1% and 10% ImageNet labels, and report both top-1 and top-5 center-crop accuracy on the ImageNet `val` set.

[19], it discusses that ImageNet pre-training shows rapid convergence than random initialization at the early stage of training but the final performance is not better than the model trained from scratch. As shown in Table 5, the scratch can narrow the gap and even surpass BYOL at last, even though the BYOL pre-training provides a slightly better initial point. Notably, UniVIP goes beyond the limitation and the noticeable gain is preserved even in longer schedules. We argue that UniVIP provides versatile representations of quality that the previous pre-training methods have not yet achieved.

4.3. Semi-supervised learning

We evaluate the performance obtained when fine-tuning UniVIP's representation on the semi-supervised learning task, following the protocol of [17, 44]. In particular, we acquire 1% and 10% labeled data from ImageNet's `train` set. Furthermore, we fine-tune our models on these two training subsets and report both top-1 and top-5 accuracies on the `val` set of ImageNet in Table 6. UniVIP consistently outperforms previous SOTA methods when pre-trained on the non-iconic and single-centric-object dataset.

4.4. Ablation studies

Effect of different levels. As shown in Table 7, it shows the effectiveness of our proposed scene, scene-instance,

	Scene	Scene-instance	Instance	Top-1	mAP
(a)	×	×	×	57.3	39.5
(b)	✓	×	×	57.7(+0.4)	39.6(+0.1)
(c)	✓	✓	×	58.1(+0.8)	40.4(+0.9)
(d)	✓	×	✓	59.6(+2.3)	40.5(+1.0)
(e)	✓	✓	✓	**60.2**(+2.9)	**40.8**(+1.3)

Table 7. **Ablations for UniVIP: Effect of different levels when pre-trained on COCO dataset.** We report linear evaluation on ImageNet and detection results on COCO.

and instance levels. The results of Table 7(a) are that we reproduce the performance of BYOL pre-training on MS COCO as the baseline. Compared with the baseline, the scene-level pre-training (Table 7(b)) can improve the performance of the neural network. Moreover, adding either the scene-instance-level (Table 7(c)) or instance-level pre-training (Table 7(d)) based on the scene level can enhance the performance, meanwhile, the best results (Table 7(e)) are acquired by adding three terms.

Effect of the scene similarity. Table 8(a) ablates the effect of the scene similarity. The "no" uses the same methods with UniVIP except for adopting two random scene views, which results in scene inconsistency. UniVIP obviously outperforms its results. Therefore, the scene similarity is crucial to self-supervised learning in natural images.

Effect of region candidates. For validating the effect of region candidates, Table 8(b) shows the results of the baseline in "none". Then, we observe that the "ground truth" of COCO can develop the performance of the pre-trained model, validating that the gains are truly due to the instance-level representation learning mechanism. Furthermore, the performance of the "naive" strategy is slightly better than the ground truth. In fact, COCO only contains manual annotations for 80 classes but the scene images have extensive unknown classes. The naive method can obtain regions where more categories are located, even if they are very crude. In this case, the diversity can make up for the inaccuracy. Nevertheless, this phenomenon does not mean that the regions containing the instance are useless. The "selective search" results indicate that the instance-based regions effectively improve the performance. The more diverse region proposal method can greatly enhance the feasibility of the model, and also evaluate that the gains are truly due to instance-level representation pre-training.

Effect of the number of instance-based views. As shown in Table 8(c), we observe that UniVIP already outperforms the state-of-the-art method [44] when the K is set as 2. The best results can be obtained when further increasing K to 4. Meanwhile, the performance slightly degrades as K increases. We argue that using 4 candidate proposals in the overlapping regions can meet most scenes, and more candidates may bring noises, thus hurting the performance.

	Scene similarity	Top-1	mAP
(a)	no	59.3	40.3
	yes	**60.2**	**40.8**
	Region candidates	Top-1	mAP
(b)	none	57.7	39.6
	ground truth	58.5	40.3
	naive	58.9	40.5
	selective search	**60.2**	**40.8**
	Number of instance views	Top-1	mAP
(c)	0	57.7	39.6
	2	59.8	40.4
	4	**60.2**	**40.8**
	8	59.7	40.6

Table 8. **Ablations for UniVIP.** (a) Effect of the scene similarity. (b) Effect of region candidates. (c) Effect of the number of instance-based views. We report linear evaluation on ImageNet and detection results on COCO.

5. Conclusion

In this paper, we analyze the two problems of current visual self-supervised learning: 1) The SSL methods with semantic consistency assumption would be infeasible for non-iconic datasets since the random views of non-iconic images are semantic inconsistency. 2) The non-iconic SSL methods hardly extract versatile visual representations. To overcome the above problems, we introduce a novel unified self-supervised learning method called UniVIP. By simultaneously leveraging the similarity of scene-scene, the correlation of scene-instance, and the discrimination of instance-instance, UniVIP can improve the versatile performance of self-supervised learning with any images. The proposed UniVIP shows versatility and scalability performance in multiple downstream visual tasks, such as image classification, semi-supervised, object detection and segmentation. We expect that our study can attract the community's attention to more versatile un-/self-supervised representation learning from natural images.

Limitations. In this work, we validate the performance of UniVIP by constructing experiments on COCO(+) dataset, and further scale our method up on ImageNet dataset. Compared with previous SSL methods, the COCO(+) and ImageNet pre-trained models all perform impressive improvements. In future, we will scale UniVIP with larger architectures and datasets [35, 36, 52] to unleash its potential.

References

[1] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Pi-otr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *Advances in Neural Information Processing Systems*, volume 33, pages 9912–9924, 2020. 1, 2, 3, 5, 6

[2] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. *arXiv: Computer Vision and Pattern Recognition*, 2021. 3

[3] Kai Chen, Lanqing Hong, Hang Xu, Zhenguo Li, and Dit-Yan Yeung. Multisiam: Self-supervised multi-instance siamese representation learning for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7546–7554, 2021. 2

[4] Mark Chen, Alec Radford, Rewon Child, Jeffrey K Wu, Hee-woo Jun, David Luan, and Ilya Sutskever. Generative pre-training from pixels. In *ICML*, volume 1, pages 1691–1703, 2020. 1

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020. 1, 2, 5, 6, 7

[6] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E. Hinton. Big self-supervised models are strong semi-supervised learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 22243–22255, 2020. 1, 2, 4

[7] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 1, 2, 5, 6, 7

[8] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15750–15758, 2021. 5

[9] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. *arXiv preprint arXiv:2104.02057*, 2021. 3

[10] Zhiyang Chen, Yousong Zhu, Chaoyang Zhao, Guosheng Hu, Wei Zeng, Jinqiao Wang, and Ming Tang. Dpt: Deformable patch-based transformer for visual recognition. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 2899–2907, 2021. 3

[11] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *NeurIPS*, 2013. 5

[12] 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. 1, 2, 3

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 3

[14] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010. 3

[15] Zheng Ge, Songtao Liu, Zeming Li, Osamu Yoshie, and Jian Sun. Ota: Optimal transport assignment for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 303–312, 2021. 2, 5

[16] Priya Goyal, Dhruv Mahajan, Abhinav Gupta, and Ishan Misra. Scaling and benchmarking self-supervised visual representation learning. *ICCV*, 2019. 5

[17] 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, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning. In *NeurIPS*, volume 33, pages 21271–21284, 2020. 1, 2, 3, 4, 5, 6, 7

[18] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. *arXiv preprint arXiv:1911.05722*, 2019. 1, 2, 4, 5, 6, 7

[19] Kaiming He, Ross Girshick, and Piotr Dollár. Rethinking imagenet pre-training. In *ICCV*, 2019. 7

[20] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017. 1, 2, 7

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1, 2, 6

[22] Cheng-I Lai. Contrastive predictive coding based feature for automatic speaker verification. *arXiv preprint arXiv:1904.01575*, 2019. 5

[23] Zhaowen Li, Zhiyang Chen, Fan Yang, Wei Li, Yousong Zhu, Chaoyang Zhao, Rui Deng, Liwei Wu, Rui Zhao, Ming Tang, and Jinqiao Wang. Mst: Masked self-supervised transformer for visual representation. In *NeurIPS*, 2021. 1, 3

[24] Tsung-Yi Lin, Piotr Dollar, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 936–944, 2017. 7

[25] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2):318–327, 2020. 1, 2

[26] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 2, 3

[27] Songtao Liu, Zeming Li, and Jian Sun. Self-emd: Self-supervised object detection without imagenet. *arXiv preprint arXiv:2011.13677*, 2020. 2, 3, 6, 7

[28] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 6

[29] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6707–6717, 2020. 2, 3, 5, 6

[30] Pedro O Pinheiro, Amjad Almahairi, Ryan Y Benmalek, Florian Golemo, and Aaron Courville. Unsupervised learning of dense visual representations. *arXiv preprint arXiv:2011.05499*, 2020. 2, 3

[31] Senthil Purushwalkam and Abhinav Gupta. Demystifying contrastive self-supervised learning: Invariances, augmentations and dataset biases. In *NIPS*, 2020. 3

[32] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*, 2015. 1, 2

[33] Byungseok Roh, Wuhyun Shin, Ildoo Kim, and Sungwoong Kim. Spatially consistent representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1144–1153, 2021. 2, 3, 6, 7

[34] Ramprasaath R Selvaraju, Karan Desai, Justin Johnson, and Nikhil Naik. Casting your model: Learning to localize improves self-supervised representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11058–11067, 2021. 2, 3

[35] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE international conference on computer vision*, pages 843–852, 2017. 3, 8

[36] 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. 2, 3, 8

[37] Yonglong Tian, Olivier J Henaff, and Aaron van den Oord. Divide and contrast: Self-supervised learning from uncurated data. In *ICCV*, 2021. 3

[38] Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders. Selective search for object recognition. *International journal of computer vision*, 104(2):154–171, 2013. 2

[39] Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Dense contrastive learning for self-supervised visual pre-training. *arXiv preprint arXiv:2011.09157*, 2020. 6, 7

[40] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019. 7

[41] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *CVPR*, 2018. 1, 5, 6

[42] Tete Xiao, Colorado J Reed, Xiaolong Wang, Kurt Keutzer, and Trevor Darrell. Region similarity representation learning. *arXiv preprint arXiv:2103.12902*, 2021. 6, 7

[43] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8392–8401, 2021. 6, 7

[44] Jiahao Xie, Xiaohang Zhan, Ziwei Liu, Yew Soon Ong, and Chen Change Loy. Unsupervised object-level representation learning from scene images. In *NeurIPS*, 2021. 3, 5, 6, 7, 8

[45] Zhenda Xie, Yutong Lin, Zhuliang Yao, Zheng Zhang, Qi Dai, Yue Cao, and Han Hu. Self-supervised learning with swin transformers. *arXiv preprint arXiv:2105.04553*, 2021. 6

[46] Ceyuan Yang, Zhirong Wu, Bolei Zhou, and Stephen Lin. Instance localization for self-supervised detection pretraining. *arXiv preprint arXiv:2102.08318*, 2021. 6, 7

[47] Yunze, GAO, Yingying, CHEN, Jinqiao, WANG, Hanqing, and LU. Progressive rectification network for irregular text recognition. *Science China(Information Sciences)*, v.63(02):7–20, 2020. 1

[48] Xiaohua Zhai, Avital Oliver, Alexander Kolesnikov, and Lucas Beyer. S4l: Self-supervised semi-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1476–1485, 2019. 7

[49] Xiao Zhang and Michael Maire. Self-supervised visual representation learning from hierarchical grouping. In *arXiv preprint arXiv:2012.03044*, 2020. 3

[50] Nanxuan Zhao, Zhirong Wu, Rynson W. H. Lau, and Stephen Lin. What makes instance discrimination good for transfer learning. In *ICLR*, 2021. 5, 7

[51] Yucheng Zhao, Guangting Wang, Chong Luo, Wenjun Zeng, and Zheng-Jun Zha. Self-supervised visual representations learning by contrastive mask prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10160–10169, 2021. 3, 6

[52] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017. 2, 3, 8