Can Semantic Labels Assist Self-Supervised Visual Representation Learning?

Longhui Wei,[1,2] **Lingxi Xie,**[2] **Jianzhong He,**[2] **Xiaopeng Zhang,**[2] **Qi Tian,**[2].

[1]University of Science and Technology of China
[2]Huawei Inc.
weilh2568@gmail.com, 198808xc@gmail.com, q-tian@hotmail.com

Abstract

Recently, contrastive learning has largely advanced the progress of unsupervised visual representation learning. Pre-trained on ImageNet, some self-supervised algorithms reported higher transfer learning performance compared to fully-supervised methods, seeming to deliver the message that human labels hardly contribute to learning transferrable visual features. In this paper, we defend the usefulness of semantic labels but point out that fully-supervised and self-supervised methods are pursuing different kinds of features. To alleviate this issue, we present a new algorithm named **Supervised Contrastive Adjustment in Neighborhood** (SCAN) that maximally prevents the semantic guidance from damaging the appearance feature embedding. In a series of downstream tasks, SCAN achieves superior performance compared to previous fully-supervised and self-supervised methods, and sometimes the gain is significant. More importantly, our study reveals that semantic labels are useful in assisting self-supervised methods, opening a new direction for the community.

Introduction

Self-supervised learning has shown its potential in various fields of artificial intelligence. In the natural language processing (NLP) community, researchers have shown a promising pipeline that uses a large amount of unlabeled data for model pre-training (*e.g.*, BERT (Devlin et al. 2018) and GPT-3 (Brown et al. 2020)) and then fine-tunes them in a series of downstream tasks. This idea is attracting increasing attentions in the computer vision community, yet the key is to define proper pretext tasks, *e.g.*, solving jigsaw puzzles (Noroozi and Favaro 2016; Wei et al. 2019; Doersch, Gupta, and Efros 2015), colorization (Zhang, Isola, and Efros 2016), to facilitate learning efficient visual representations from unlabeled image data.

Recently, this progress is largely advanced by contrastive learning (He et al. 2020; Chen et al. 2020a,b; Tian, Krishnan, and Isola 2019; Tian et al. 2020; Grill et al. 2020; Krishna et al. 2020). The assumption is that each image can be encoded into a low-dimensional feature vector and the encoding scheme is resistant against data augmentation (*e.g.*, image cropping and horizontal flipping) so that the algorithm

Dataset	Task	Metric	Pre-Training Method		
			FSup	MoCo	**SCAN**
VOC	Det	AP_{50}	81.4	82.5	**83.4**
	Seg	mIoU	74.4	73.4	**76.6**
COCO	Det	AP^{bb}	38.9	39.2	**40.9**
	InsSeg	AP^{mk}	35.4	35.7	**37.2**
Cityscapes	InsSeg	AP	32.9	33.1	**33.8**
	Seg	mIoU	74.6	75.2	**76.5**

Table 1: Comparison among three pre-training methods, namely, fully-supervised (FSup), MoCo (v2 (Chen et al. 2020b)), and SCAN (our approach), in the performance (%) of downstream transfer. ResNet-50 is used as the standard backbone, and all models are pre-trained only on ImageNet-1K. SCAN enjoys advantages over a wide range of datasets, tasks, and evaluation protocols. Please refer to the experimental section for full results.

can distinguish the instance from a large gallery containing its variant as well as other instances. Contrastive learning has claimed significant improvement over traditional pretext tasks and, interestingly, the performance in some popular downstream tasks (including object detection and image segmentation, see Tab. 1) has surpassed the fully-supervised counterpart. This raises an important question: are the semantic labels useless for transferring the learned visual representations to other scenarios?

This paper aims to answer the above question and defend the usefulness of semantic labels in visual representation learning. We first point out that fully-supervised and self-supervised methods are trying to capture different kinds of features for visual representation. We refer to them as **task-specific semantic features** and **task-agnostic appearance features**, respectively, where the former mainly focuses on semantic understanding and the latter is better at instance discrimination. Typical examples are shown in Fig. 1 where we extract features from a query image and try to find its nearest neighbors in the gallery. When the semantic features are used, the returned images are mostly from the same object class but the general layout may be quite different; when the appearance features are used, the opposite phenomenon is observed. Though the appearance features are verified to have better transfer ability, we expect the semantic features to assist visual understanding, *e.g.*, improving the classifica-
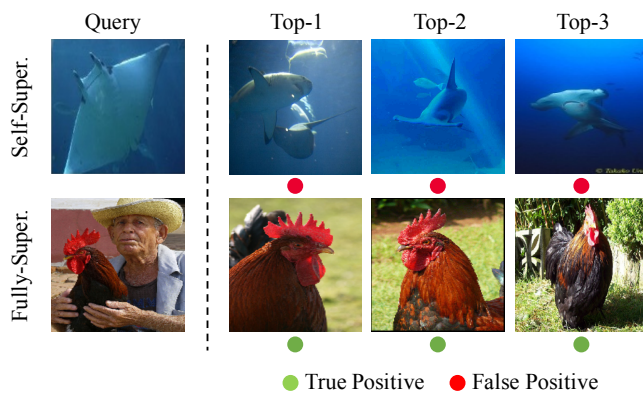
Figure 1: Examples of retrieval results with the features generated by self-supervised (MoCo-v2) and traditional fully-supervised method (trained with Softmax loss). In each example, the top-3 most similar images are displayed. True Positive means that the returned images belong to the same class of the query, and False Positive implies different classes. Clearly, self-supervised learning pays more attention to global appearance than local details (though the details sometimes imply class information), but fully-supervised learning learns to focus on the semantic regions.

tion scores in detection and segmentation. This motivates us to design an algorithm that absorbs the advantages of both kinds of features.

However, we shall be aware that if fully-supervised and self-supervised objectives are directly combined, they may conflict with each other and thus deteriorate the performance of downstream tasks. To avoid the conflict, we present a novel algorithm named **Supervised Contrastive Adjustment in Neighborhood** (SCAN). The core part is a new definition of image-level relationship. Note that fully-supervised learning assumes that all images from the same class shall produce similar feature representations, while self-supervised contrastive learning aims to distinguish all different images. SCAN is kind of compromise between them, trying to **enhance the relationship between each instance and others that are close to it in both semantics and appearance.**

In practice, SCAN is implemented by inheriting a pre-trained model by unsupervised contrastive learning, *e.g.*, MoCo (Chen et al. 2020b). Then, for each query, instead of treating all other images in the gallery as negative samples, SCAN preserves several positive samples that are similar to the query in both semantics and appearance. The contrastive learning objective is hence modified so that positive samples are pulled towards each other while the negative samples are pushed away as usual. The required efforts in code editing beyond the existing contrastive learning methods are minor.

We evaluate SCAN on a wide range of downstream tasks for detection and segmentation. As shown in Tab. 1, SCAN outperforms the fully-supervised and self-supervised baselines consistently, sometimes significantly. Moreover, though the state-of-the-art self-supervised learning methods (Chen et al. 2020a; He et al. 2020) claimed the benefits

from extensive (around 1 billion) unlabeled images, SCAN is able to surpass their performance by training on labeled ImageNet-1K with around 1 million images. More experimental results have shown in Sec. .

Related Work

Our work is proposed under the current state-of-the-art self-supervised and fully-supervised learning methods. In this section, we mainly summarize the existing methods of the above aspects and elaborate the differences between these methods and our approach.

Self-Supervised Learning Methods

The key to solving the self-supervised learning is how to design a pretext task. In the early stage, some novel hand-crafted pretext tasks were proposed to extract useful knowledge. For example, some works (Vincent et al. 2008; Pathak et al. 2016; Zhang, Isola, and Efros 2016, 2017; Kingma and Welling 2013; Rezende, Mohamed, and Wierstra 2014) designed the image restoration related pretext tasks, and utilized auto-encoding based approaches to recover the input image. Other pretext tasks, *e.g.*, jigsaw puzzle (Noroozi and Favaro 2016; Wei et al. 2019; Doersch, Gupta, and Efros 2015), image rotation prediction (Gidaris, Singh, and Komodakis 2018; Chen et al. 2019), *etc*, have also been proposed to further push forward the final performance.

Recently, instance discrimination related methods have achieved great success in the self-supervised learning field (Wu et al. 2018; Gutmann and Hyvärinen 2010; Zhuang, Zhai, and Yamins 2019; Misra and Maaten 2020; Oord, Li, and Vinyals 2018; Hénaff et al. 2019; Tian, Krishnan, and Isola 2019; Hjelm et al. 2018; Li et al. 2021). Non-parametric Instance Discrimination (Wu et al. 2018) utilized the novel instance-level discrimination idea to extract useful knowledge from extensive unlabeled images. Contrastive Multiview Coding (Tian, Krishnan, and Isola 2019) successfully extended the two views to more views of contrastive learning. Momentum Contrast (MoCo) (He et al. 2020) built a dynamic but representation-consistent memory bank for contrastive learning. SimCLR (Chen et al. 2020a) evaluated that larger batch size and suitable composition of data augmentation operations have clear benefits to the current contrastive learning framework. Prototypical Contrastive Learning (Li et al. 2021) and SwAV (Caron et al. 2020) proposed online clustering-based methods to learn more discriminative representation. BYOL (Grill et al. 2020) firstly evaluated the self-supervised learning could still work well without utilizing negative pairs. Moreover, Dense Contrastive Learning (DenseCL) (Wang et al. 2021) and Pixel-to-Propagation (PixPro) (Xie et al. 2021) further enhanced the representation ability in pixel-level.

The appealing points for self-supervised learning not only lie on they can extract useful knowledge from unlabeled data, but also they have evaluated that even trained on the same data, they still benefit more to downstream tasks compared with previous fully-supervised learning methods. Differently, our method targets to reveal the opposite conclusion that, with rationally leveraging the human annotations, we
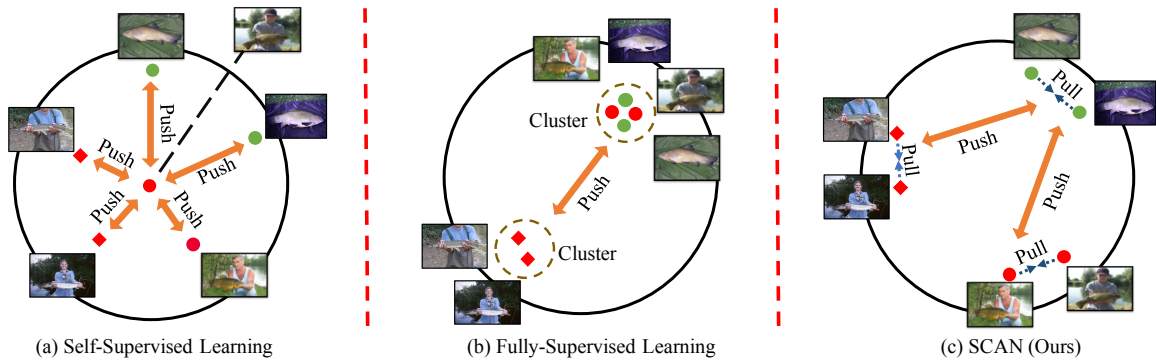
Figure 2: Illustrations of representation learning for self-supervised learning (instance discrimination based), fully-supervised learning, and our approach. Different icon shapes denote different classes, and different icon colors represent the corresponding images are dissimilar in global appearance. Different from self-supervised learning aiming to achieve instance-level discrimination and fully-supervised learning targeting to cluster the images of the same annotated class, SCAN tries to pull appearance-similar images of the same class closer and push other images away.

can achieve much better performance of pre-training models on downstream tasks.

Fully-Supervised Learning Methods

For fully-supervised learning, the two most related aspects to our work are data augmentation methods and loss functions, respectively. Multiple novel data augmentation methods have been proposed in previous years, *e.g.*, Cutout (De-Vries and Taylor 2017), Mixup (Zhang et al. 2017) and AutoAugment (Cubuk et al. 2018, 2020). As for loss functions, Softmax loss is the most commonly used loss function. Moreover, label smoothing (Müller, Kornblith, and Hinton 2019) and knowledge distillation related loss functions (Hinton, Vinyals, and Dean 2015; Wei et al. 2020) are also widely used as the auxiliary loss. Recently, Supervised Contrastive Learning (SCL) has successfully extended the contrastive loss (Khosla et al. 2020) from self-supervised learning field to fully-supervised learning field, and it achieved competitive performance compared with the traditional Softmax loss. Though fully-supervised learning methods have achieved excellent performance on target tasks, the learned semantic features are biased and cannot transfer well into downstream tasks. Differently, our approach targets to learn a general pre-training feature space with the guidance of semantic labels and self-supervised learning methods.

Our Method

Preliminaries: Momentum Contrast

Momentum Contrast (MoCo) is a current state-of-the-art self-supervised learning method. Similar to most of the previous methods (Tian, Krishnan, and Isola 2019; Wu et al. 2018; Chen et al. 2020a), MoCo also chooses the instance discrimination mechanism as its pretext task. Differently, MoCo builds a memory bank for enlarging the look-up dictionary, and maintains an additional encoder by momentum update policy to dynamically update features in the memory bank. Through the above strategy, MoCo can nearly ignore the limitation of GPU memory, and compare a query sample

with countless queue samples in the memory bank. Therefore, suppose a query image as $\mathbf{q}_i$, its two corresponding augmented images are denoted as $\mathbf{q}_i^{\dagger}$ and $\mathbf{q}_i^{\ddagger}$, the encoders for the query and memory bank are $\mathbf{f}$ and $\mathbf{g}$, the memory bank size is L, and the feature of each sample embedded in this memory bank is represented as $\mathbf{z}_l$, the loss function of MoCo for this query image can be formulated as:

$$\mathcal{L}_i^{\text{MoCo}} = -\log \frac{\exp\{\mathbf{f}(\mathbf{q}_i^{\dagger})^{\top} \cdot \mathbf{g}(\mathbf{q}_i^{\ddagger})/\tau\}}{\text{Sim}_i^{\text{all}}}, \qquad (1)$$

in which

$$\text{Sim}_i^{\text{all}} = \exp\{\mathbf{f}(\mathbf{q}_i^{\dagger})^{\top} \cdot \mathbf{g}(\mathbf{q}_i^{\ddagger})/\tau\}$$
$$+ \sum_{l=1}^{L} \exp\{\mathbf{f}(\mathbf{q}_i^{\dagger})^{\top} \cdot \mathbf{z}_l/\tau\} \qquad (2)$$

where τ represents the temperature hyper-parameter.

As shown in Eq. (1), the goal of MoCo is to pull in the same instance with different transformations, and enlarge the distances among the query and samples in the memory bank. With this simple but effective scheme, MoCo heavily advances the development of self-supervised learning field.

Conflict between Fully-Supervised and Self-Supervised Learning

Though MoCo has evaluated self-supervised learning methods can learn more transferrable knowledge, it is still doubtful that extra human annotations show no advantages for pre-training models. Compared with the raw images, human annotations can be regarded as extra semantic information. Therefore, the labeled dataset provides more knowledge than the unlabeled data, and in theory, the learned knowledge by self-supervised learning methods should be one subset of that learned by fully-supervised learning methods. However, the existing experiments reveal an opposite conclusion that using semantic labels harms the downstream performance. To investigate this paradox, we delve deep into

the representation learning natures in the self-supervised and fully-supervised learning setting.

Interestingly, we find self-supervised and fully-supervised learning are pursuing different kinds of features, **task-agnostic appearance feature** and **task-specific semantic feature**. As shown in Fig. 2, for each query image, these self-supervised learning methods based on contrastive learning (like MoCo) target to push away all different samples for achieving the goal of instance-level discrimination. Consequently, they tend to extract the global appearance feature for roughly describing the whole image. In their feature spaces, the appearance-similar images will locate closer and the appearance-dissimilar images will be far away. Therefore, the learned knowledge by self-supervised approaches is nearly task-agnostic, and can be well transferred into downstream tasks. However, as shown in Fig. 1, the generated appearance feature hardly describe the semantic details, which reflects there is still much room to improve. Differently, fully-supervised learning methods pull images of the same class into one cluster and push images of the different classes away, ignoring other cues. Hence, fully-supervised algorithms will always attend on the limited task-relevant regions. Though the final generated semantic feature is more discriminative, it is heavily biased and cannot ensure the performance on every downstream task.

To summarize, the representation learning goal of self-supervised and fully-supervised learning methods are totally different, and there are both advantages and weaknesses of them for visual pre-training task. Therefore, how to relieve the conflict of representation learning on them and effectively exploit their unique superiority is a good direction to improve the pre-training models. We shall present our solution in the next part.

Our Solution: Supervised Contrastive Adjustment in Neighborhood

The aforementioned conflict mainly comes from the inconsistency between the semantic and appearance spaces – some samples are similar in semantics but not in appearance, and the opposite situation can also happen. To maximally avoid it, we design a conservative method that only takes the intersection of the semantic and appearance neighborhoods into consideration. Our algorithm is named **Supervised Contrastive Adjustment in Neighborhood** (SCAN). As shown in Fig. 2, SCAN is surprisingly simple: an extra judgment is added beyond the semantic labels, filtering the images sharing the same class label but the appearance is not sufficiently close to the query. Then, these remained images similar in both semantics and appearance to the query are pulled together, and other images are pushed away. In this way, SCAN maximally keeps the consistency between semantics and appearance similarity, and learns both kinds of knowledge simultaneously.

Therefore, the most important task for SCAN is to find the images that are similar to each other in both semantics and appearance. Mathematically, for each image pair, $\mathbf{x}_i$ and $\mathbf{x}_j$, the semantic similarity of them can be simply measured by

the corresponding semantic labels, as follows:

$$\text{Sim}_{\text{S}}(\mathbf{x}_i, \mathbf{x}_j) = 1_{\mathbf{y}_i = \mathbf{y}_j}, \tag{3}$$

where 1 represents the indicator function, and $\mathbf{y}_i$ denotes the class label of $\mathbf{x}_i$. Since there is no ground-truth for measuring the appearance similarity, we refer to any pre-trained models that have not used semantic labels (*e.g.*, MoCo-v2) for feature extraction. Let $\mathbf{a}_i$ and $\mathbf{a}_j$ be the features of $\mathbf{x}_i$ and $\mathbf{x}_j$, the appearance similarity of every two images is computed by:

$$\text{Sim}_{\text{A}}(\mathbf{x}_i, \mathbf{x}_j) = \frac{1}{2}\left(\mathbf{a}_i^{\top} \cdot \mathbf{a}_j + 1\right), \tag{4}$$

Note that the values of both $\text{Sim}_{\text{S}}(\cdot)$ and $\text{Sim}_{\text{A}}(\cdot)$ falls in the range of $[0, 1]$, with larger values indicating higher similarity. Integrating Eq. (3) and Eq. (4) yields the similarity measurement that takes both semantics and appearance into consideration:

$$\text{Sim}_{\text{S\&A}}(\mathbf{x}_i, \mathbf{x}_j) = \text{Sim}_{\text{S}}(\mathbf{x}_i, \mathbf{x}_j) \times \text{Sim}_{\text{A}}(\mathbf{x}_i, \mathbf{x}_j), \tag{5}$$

That being said, two images are considered similar only if they share similar layouts and the same semantic label.

Eq. (5) offers a new way to generate the pseudo labels for contrastive learning. Given the query image, we compute its similarity to each candidate in the gallery using Eq. (5), and retrieve the top-ranked images with non-zero similarity as the positive samples (or called positive neighbors). This avoids the burden that two images with either distinct semantic or appearance features to be assigned the identical pseudo label, and thus alleviates the risk of conflict in the feature space.

Naturally, the generated pseudo labels embed well both the appearance and semantic information. Through simply pushing away images of the different assigned pseudo classes, SCAN inherits the ability of capturing appearance features. By pulling images of the same pseudo class together, SCAN is forced to embed semantic information into the feature space. In our implementation, limited by the GPU memory, SCAN also resorts to maintain a memory bank as MoCo for storing a large scale of samples appearing in the previous mini-batch, which will be regarded as negative samples in the current mini-batch. Therefore, suppose a total of S images and their corresponding top-K positive neighbors are sampled in one mini-batch, the corresponding loss function of SCAN for each query image $\mathbf{x}_i$ can be formulated as:

$$\mathcal{L}_i^{\text{SCAN}} = -\frac{1}{K} \sum_{j=1}^{S(K+1)} 1_{\mathbf{y}_i^* = \mathbf{y}_j^*}$$
$$- \log \frac{\exp\{\mathbf{f}(\mathbf{x}_i^{\dagger})^{\top} \cdot \mathbf{g}(\mathbf{x}_j^{\ddagger})/\tau\}}{\text{Sim}_i^{\text{all}}}, \tag{6}$$

in which

$$\text{Sim}_i^{\text{all}} = \sum_{t=1}^{S(K+1)} \exp\{\mathbf{f}(\mathbf{x}_i^{\dagger})^{\top} \cdot \mathbf{g}(\mathbf{x}_t^{\ddagger})/\tau\}$$
$$+ \sum_{l=1}^{L} \exp\{\mathbf{f}(\mathbf{x}_i^{\dagger})^{\top} \cdot \mathbf{z}_l/\tau\} \tag{7}$$

where $\mathbf{y}_i^*$ denotes the generated pseudo label of $\mathbf{x}_i$.

It is worth noting that SCAN approximately degenerates to prior approaches, MoCo (He et al. 2020) and SCL (Khosla et al. 2020), by setting the variable of K in Eq. (6) to be 0 and ∞, respectively. Though the formulations of these three methods are similar, the motivations of them are different. Unlike MoCo targeting to learn more useful knowledge from unlabeled data, or SCL aiming for achieving better performance on the image classification task, the core of SCAN is to improve the performance of visual pre-training models. Therefore, SCAN is designed to relieve the representation learning conflict between self-supervised learning and fully-supervised learning, and then maximally learns transferrable knowledge from both appearance and semantics. We refer the readers to Sec. for further empirical analysis.

Difference from Recent Work

Some concurrent works (Kang et al. 2020; Zhao et al. 2020) seem similar to SCAN, which also inject semantic labels into contrastive learning framework for improving the transfer performance, but there are still differences from ours. For example, KCL (Kang et al. 2020) randomly samples the K positive images to tackle with long-tailed recognition. Differently, SCAN aims to relieve the conflict between self-supervised and fully-supervised learning, and thus the K positive images are obtained from similarity ranking. As shown in Tab. 2, SCAN achieves much better transfer performance compared with KCL on downstream tasks. Moreover, Exemplar-v2 (Zhao et al. 2020) has evaluated the bad effectiveness (the biased invariance) of pulling together the positives, and thus just pushes away the true negatives to improve the transfer performance. Differently, SCAN only pulls together the top-K similar positives, which can enhance the semantics and meanwhile maximally avoid to learn the biased invariance. Experimental results on Tab. 3 show SCAN can heavily improve the final transfer performance.

Experiments

Datasets and Implementation Details

Datasets. We pre-train SCAN on ImageNet-1K (Deng et al. 2009), the most widely used large-scale classification dataset. As for the downstream evaluation stage, we mainly adopt three commonly used detection and segmentation datasets, *i.e.*, PASCAL VOC (Everingham et al. 2010), COCO (Lin et al. 2014) and Cityscapes (Cordts et al. 2016), respectively. In the following of this paper, we refer to PASCAL VOC as VOC for brevity.

Implementation. We directly utilize the official released MoCo-v2 model trained on ImageNet-1K with 800 epochs to generate the appearance feature, and then compute the similarity of samples as Eq. (5) to retrieve positive neighbors accordingly. Empirically, we assign each sample and its top-2 positive neighbors to the same new label. Finally, we select ResNet-50 (He et al. 2016) as the used backbone and train it with the guidance of our generated labels. For one mini-batch of each GPU, we randomly choose 128 samples and their corresponding positive neighbors. Therefore, there

Methods	Data	VOC$_{\text{Det}}$			VOC$_{\text{Seg}}$
		AP$_{50}$	AP$_{75}$	AP	mIoU
FSup-Softmax	IN	81.4	58.8	53.5	74.4
KCL	IN	82.3	62.1	55.5	-
MoCo-v1	IG	82.2	63.7	57.2	73.6
MoCo-v2	IN	82.5	64.0	57.4	73.4
SwAV	IN	82.6	62.7	56.1	-
DenseCL	IN	82.8	**65.2**	**58.7**	-
SCAN	IN	**83.4**	64.2	57.5	**76.6**

Table 2: Results of different pre-training methods on VOC. VOC$_{\text{Det}}$/VOC$_{\text{Seg}}$ means detection and semantic segmentation task on VOC, respectively. IN denotes the pre-training dataset is ImageNet-1K, and IG represents the pre-training dataset is Instagram. FSup-Softmax means the backbone is trained with the Softmax loss under the fully-supervised setting. The reported numbers of SCAN are the average of 3 runs.

are always positive samples for each anchor image in the mini-batch. Similar to MoCo, we adopt the SGD optimizer and set momentum as 0.9. Moreover, the cosine learning rate schedule is utilized, and the initial learning rate is set as 1.6. The temperature τ in Eq. (6) is empirically set as 0.07. For the data augmentation, we simply follow the augmentation scheme in MoCo-v2. Finally, we use 32 Tesla-V100 GPUs to train our model lasting for 400 epochs on ImageNet-1K.

Results on PASCAL VOC

For the detection task on VOC, we use Detectron2 to fine-tune the Faster-RCNN (Ren et al. 2015) with R50-C4 backbone. As for the semantic segmentation task, we use the same FCN-based structure (Long, Shelhamer, and Darrell 2015) as MoCo to conduct the evaluations. We strictly follow the training and testing settings in MoCo, and the only difference is that the training iterations are extended from 30K to 50K in semantic segmentation task (to successfully reproduce the reported result in MoCo).

The detection and semantic segmentation results on VOC are shown in Tab. 2. One interesting phenomenon is that, self-supervised and fully-supervised learning methods show different properties while transferring their learned knowledge into different downstream tasks. In the detection task, MoCo-v2 surpasses FSup-Softmax with a large margin, *e.g.*, 3.9% improvement on Average Precision (AP) metric. However, there is 1.0% performance degradation in the semantic segmentation task. Generally, the above results further reveal that, there exist both advantages and weaknesses of self-supervised and fully-supervised learning methods for the visual pre-training models.

Benefiting from effectively exploiting their unique advantages of self-supervised and fully-supervised learning methods, SCAN heavily pushes forward the final transfer performances of pre-training models. For example, compared with the baseline (MoCo-v2), there is 0.9% improvement on AP$_{50}$ metric for SCAN wile transferring the pre-trained knowledge into detection task in VOC. Moreover, for the semantic segmentation results on VOC, SCAN significantly improves the transfer performance of MoCo-v2 from **73.4%**

| Methods | Data | Mask R-CNN, R50-FPN, COCO$_{Det}$ | | | | | | Mask R-CNN, R50-FPN, COCO$_{InsSeg}$ | | | | | |
| | | 1× schedule | | | 2× schedule | | | 1× schedule | | | 2× schedule | | |
		AP^{bb}	AP$_{50}^{bb}$	AP$_{75}^{bb}$	AP^{bb}	AP$_{50}^{bb}$	AP$_{75}^{bb}$	AP^{mk}	AP$_{50}^{mk}$	AP$_{75}^{mk}$	AP^{mk}	AP$_{50}^{mk}$	AP$_{75}^{mk}$
FSup-Softmax	IN	38.9	59.6	42.7	40.6	61.3	44.4	35.4	56.5	38.1	36.8	58.1	39.5
MoCo-v1	IG	38.9	59.4	42.3	41.1	61.8	45.1	35.4	56.5	37.9	37.4	59.1	40.2
MoCo-v2	IN	39.2	59.9	42.7	41.6	62.1	45.6	35.7	56.8	38.1	37.7	59.3	40.6
Exemplar-v2	IN	39.4	59.1	42.7	-	-	-	34.4	55.9	36.5	-	-	-
InfoMin Aug.$^\sharp$	IN	40.6	60.6	44.6	42.5	62.7	46.8	36.7	57.7	39.4	38.4	59.7	41.4
DenseCL$^\sharp$	IN	40.3	59.9	44.3	-	-	-	36.4	57.0	39.2	-	-	-
SCAN	IN	**40.9**	**61.9**	**44.7**	42.3	63.1	46.3	**37.2**	**58.8**	**39.9**	38.2	60.2	41.1
SCAN$^\sharp$	IN	**41.8**	**62.1**	**45.7**	43.2	63.3	47.1	**37.8**	**59.2**	**40.9**	38.8	60.6	41.6

Table 3: Results of utilizing different pre-training models to fine-tune Mask R-CNN with the R50-FPN backbone on COCO. COCO$_{Det}$ and COCO$_{InsSeg}$ denote the detection and instance segmentation task on COCO, respectively. $\sharp$ represents that the method follows the fine-tuning protocol in InfoMin (instead of MoCo) for Mask R-CNN with the R50-FPN backbone.

to **76.6%**. The consistent improvements in different tasks sufficiently evaluate the effectiveness of our method. It is worth noting that, because of exploring more fine-grained information in pixel-level, DenseCL can achieve much better performance on VOC compared with SCAN. However, this advantage is no longer exists on COCO and Cityscapes. More comparisons are shown in Tab. 3 and Tab. 4.

Results on MS-COCO

COCO is the most popular large-scale detection and segmentation dataset. In this section, we utilize Mask R-CNN (He et al. 2017) with the R50-FPN backbone to conduct detection and instance segmentation task concurrently. All of the experiments are trained on COCO$_{train2017}$ and tested on COCO$_{val2017}$.

As shown in Tab. 3, compared with FSup-Softmax, MoCo-v2 shows clear advantages in both detection and instance segmentation tasks, which seems to reveal that human annotations do not contribute to learning transferrable knowledge. However, with properly exploiting the available labels, SCAN can significantly outperform MoCo-v2, *e.g.*, 1.7% improvement on AP metric with 1× schedule in COCO detection task. Compared with the recent state-of-the-art methods, *e.g.*, InfoMin (Tian et al. 2020) and DenseCL (Wang et al. 2021), SCAN still shows competitive performance. **The consistent improvements on COCO demonstrate the semantic labels do benefit pre-training models if they are exploited properly**.

Results on Cityscapes

Different from VOC and COCO that consist of natural images, Cityscapes is an urban street scene dataset. Therefore, it can better evaluate the robustness of different pre-training models trained on ImageNet-1K (consisting of natural images). For the instance segmentation task in Cityscapes, we adopt Mask R-CNN with the R50-FPN backbone and 2× schedule for training. As for the semantic segmentation task, we use the same FCN-based model as MoCo to conduct the evaluations. All of the experimental settings are the same with MoCo, except that in the testing stage, we utilize slide inference policy as PSPNet (Zhao et al. 2017) for successfully reproducing the reported performance in MoCo. As shown in Tab. 4, SCAN consistently outperforms MoCo-v2

| Methods | Data | CS$_{InsSeg}$ | | CS$_{Seg}$ |
		AP	AP$_{50}$	mIoU
FSup-Softmax	IN	32.9	59.6	74.6
MoCo-v1	IG	32.9	60.3	75.5
MoCo-v2	IN	33.1	60.1	75.2
DenseCL	IN	-	-	75.7
SCAN	IN	**33.8**	**61.3**	**76.5**

Table 4: Results of different pre-training methods on instance and semantic segmentation task in Cityscapes. CS$_{InsSeg}$ and CS$_{Seg}$ represents the instance segmentation and semantic segmentation task in Cityscapes, respectively. The reported results of SCAN are the average of 3 runs.

on Cityscapes, *e.g.*, 0.7% improvement on AP metric in instance segmentation task and 1.3% improvement on semantic segmentation task, respectively.

Ablation Study

The Impact of Hyper-parameters. As introduced in Sec. , the only hyper-parameter of our approach is the candidate top-K positive neighbors. While setting the value of K to be 0, SCAN degenerates into a self-supervised learning formulation (MoCo). While seting the value of K to be ∞, SCAN changes into a fully-supervised learning formulation (SCL). Therefore, selecting a suitable value of K is important for SCAN. As shown in Fig. 3, with enlarging K, the performances of SCAN on downstream tasks are firstly improved and then slowly declined. For example, with increasing the value of K from 0 to 2, the detection performance of the pre-training model trained by SCAN is improved from 39.2% to 40.9% on COCO. That indicates enhancing the semantic discrimination ability is indeed required for pre-training models. However, while enlarging the value of K from 2 to 8, the instance segmentation performance on COCO is decreased from 37.2% to 36.3%, which reveals clustering appearance-dissimilar images of the same class will destroy the generalization ability.

Analysis of Representation Learning in SCAN. Though SCAN achieves excellent performance on different downstream tasks, there remains a question of whether it learns more useful knowledge compared with the self-supervised
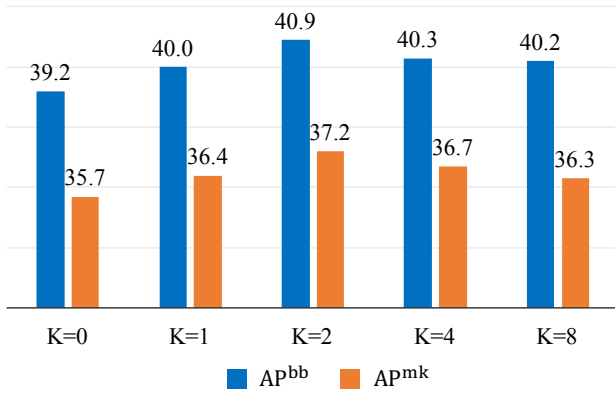
Figure 3: The impact of different values of K. AP^{bb} and AP^{mk} represents the Average Precision metric for detection and instance segmentation task on COCO by fine-tuning Mask R-CNN with the R50-FPN backbone and $1\times$ schedule, respectively.
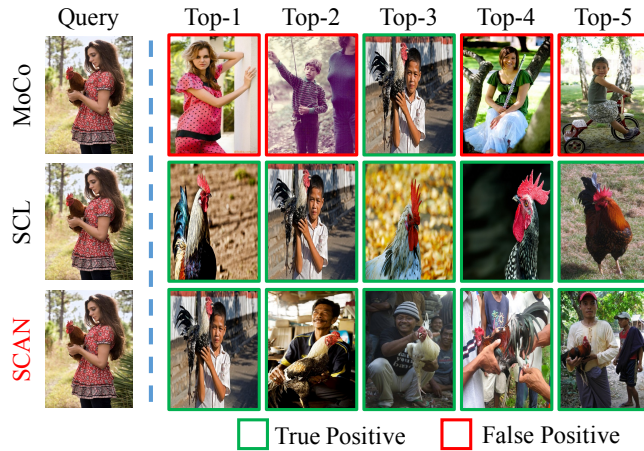

Figure 4: Examples of retrieval results with the features generated by MoCo(v2), SCL, and SCAN. True Positive means that the returned images belong to the same class of the query, and False Positive implies different classes.

and fully-supervised learning methods. To evaluate this, we conduct **visualization analysis** on ImageNet-1K. The retrieval results with the features generated by MoCo, SCL, and SCAN are shown in Fig. 4. Generally, MoCo mainly focuses on the overall layout (a person at the center of the image) but SCL fits the semantic labels (the rooster). Both of them are prone to lose part of image information. In comparison, SCAN achieves a balance between them and retrieves the images that are similar to the query in both semantics and appearance. We argue that this property is helpful in deploying the pre-trained model to the downstream tasks.

Additionally, to **quantitatively analyze** how SCAN enhances the semantic discrimination ability by pulling the positive neighbors, we conduct comparisons with recent self-supervised methods in linear evaluation task on ImageNet-1K. As shown in Tab. 5, SCAN achieves 75.8%

Methods	Top-1 Acc.(%)	Top-5 Acc.(%)
MoCo-v2	71.1	-
InfoMin Aug.	73.0	91.1
BYOL	74.3	91.6
SwAV	75.3	-
SCAN*	**75.8**	**92.4**

Table 5: Accuracy of different pre-training methods under linear evaluation on ImageNet-1K with ResNet-50 backbone. * represents the semantic labels are used.

Methods	Data	AP^{bb}_S	AP^{bb}_M	AP^{bb}_L
MoCo-v2	IN	23.8	45.6	**56.5**
SCAN	IN	**24.3**	**46.7**	56.3

Table 6: Analysis of utilizing different pre-training models to fine-tune Mask R-CNN with the R50-C4 backbone and $2\times$ schedule on the COCO detection task. AP^{bb}_S, AP^{bb}_M, and AP^{bb}_L represent the AP metric on detecting the small, middle, and large objects according to the COCO standard.

on Top-1 accuracy, with 4.7% improvement compared with MoCo-v2. Moreover, compared with other state-of-the-art methods, SCAN still shows competitive performance.

For that SCAN leverages additional labels compared with self-supervised learning methods, **we acknowledge that the above comparison experiment on ImageNet-1K is somewhat unfair**. To better evaluate that SCAN indeed learns more transferrable semantic knowledge, we further do analysis on downstream tasks. For the detection task in COCO, researchers have divided the annotated objects into three types according to their object size, *i.e.*, the small, middle, and large objects, respectively. Therefore, while detecting the large objects, the requirement of the semantic discrimination ability is relatively weaker than that while detecting the small and middle objects. Consequently, as shown in Tab. 6, SCAN indeed achieves much better performance in detecting the small and middle objects, but slightly worse performance in detecting the large objects compared with MoCo-v2. These results verify that SCAN improves visual representations by integrating the semantic information into the appearance features, and this mainly owes to the semantic labels properly leveraged.

Conclusions

In this paper, we present Supervised Contrastive Adjustment in Neighborhood (SCAN), the first work showing that making good use of semantic labels in contrastive learning can benefit downstream tasks. By simply pulling each query image and its *positive neighbors* together while pushing other images away, SCAN can maximally learn knowledge from both appearance and semantics. Extensive experiments on various downstream tasks clearly demonstrated the effectiveness of SCAN. Importantly, our research reveals the potential of making *proper* use of supervision labels in assisting the self-supervised visual representation learning and we believe this direction is important for the community.

References

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Caron, M.; Misra, I.; Mairal, J.; Goyal, P.; Bojanowski, P.; and Joulin, A. 2020. Unsupervised learning of visual features by contrasting cluster assignments. *arXiv preprint arXiv:2006.09882*.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020a. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*.

Chen, T.; Zhai, X.; Ritter, M.; Lucic, M.; and Houlsby, N. 2019. Self-supervised gans via auxiliary rotation loss. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 12154–12163.

Chen, X.; Fan, H.; Girshick, R.; and He, K. 2020b. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*.

Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; and Schiele, B. 2016. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3213–3223.

Cubuk, E. D.; Zoph, B.; Mane, D.; Vasudevan, V.; and Le, Q. V. 2018. Autoaugment: Learning augmentation policies from data. *arXiv preprint arXiv:1805.09501*.

Cubuk, E. D.; Zoph, B.; Shlens, J.; and Le, Q. V. 2020. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 702–703.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. ImageNet: A large-scale hierarchical image database. In *CVPR*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

DeVries, T.; and Taylor, G. W. 2017. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*.

Doersch, C.; Gupta, A.; and Efros, A. A. 2015. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE international conference on computer vision*, 1422–1430.

Everingham, M.; Van Gool, L.; Williams, C. K.; Winn, J.; and Zisserman, A. 2010. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2): 303–338.

Gidaris, S.; Singh, P.; and Komodakis, N. 2018. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*.

Grill, J.-B.; Strub, F.; Altché, F.; Tallec, C.; Richemond, P. H.; Buchatskaya, E.; Doersch, C.; Pires, B. A.; Guo, Z. D.; Azar, M. G.; et al. 2020. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*.

Gutmann, M.; and Hyvärinen, A. 2010. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, 297–304.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9729–9738.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, 2961–2969.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hénaff, O. J.; Srinivas, A.; De Fauw, J.; Razavi, A.; Doersch, C.; Eslami, S.; and Oord, A. v. d. 2019. Data-efficient image recognition with contrastive predictive coding. *arXiv preprint arXiv:1905.09272*.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Hjelm, R. D.; Fedorov, A.; Lavoie-Marchildon, S.; Grewal, K.; Bachman, P.; Trischler, A.; and Bengio, Y. 2018. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*.

Kang, B.; Li, Y.; Xie, S.; Yuan, Z.; and Feng, J. 2020. Exploring balanced feature spaces for representation learning. In *International Conference on Learning Representations*.

Khosla, P.; Teterwak, P.; Wang, C.; Sarna, A.; Tian, Y.; Isola, P.; Maschinot, A.; Liu, C.; and Krishnan, D. 2020. Supervised contrastive learning. *arXiv preprint arXiv:2004.11362*.

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Krishna, C.; Ertunc, E.; Neerav, K.; and Ender, K. 2020. Contrastive learning of global and local features for medical image segmentation with limited annotations. *arXiv preprint arXiv:2006.10511*.

Li, J.; Zhou, P.; Xiong, C.; Socher, R.; and Hoi, S. C. 2021. Procotypical contrastive learning of unsupervised representation. In *International Conference on Learning Representation*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, 740–755. Springer.

Long, J.; Shelhamer, E.; and Darrell, T. 2015. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3431–3440.

Misra, I.; and Maaten, L. v. d. 2020. Self-supervised learning of pretext-invariant representations. In *Proceedings of*

the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 6707–6717.

Müller, R.; Kornblith, S.; and Hinton, G. E. 2019. When does label smoothing help? In *Advances in Neural Information Processing Systems*, 4694–4703.

Noroozi, M.; and Favaro, P. 2016. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European Conference on Computer Vision*, 69–84. Springer.

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

Pathak, D.; Krahenbuhl, P.; Donahue, J.; Darrell, T.; and Efros, A. A. 2016. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2536–2544.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, 91–99.

Rezende, D. J.; Mohamed, S.; and Wierstra, D. 2014. Stochastic backpropagation and variational inference in deep latent gaussian models. In *International Conference on Machine Learning*, volume 2.

Tian, Y.; Krishnan, D.; and Isola, P. 2019. Contrastive multiview coding. *arXiv preprint arXiv:1906.05849*.

Tian, Y.; Sun, C.; Poole, B.; Krishnan, D.; Schmid, C.; and Isola, P. 2020. What makes for good views for contrastive learning. *arXiv preprint arXiv:2005.10243*.

Vincent, P.; Larochelle, H.; Bengio, Y.; and Manzagol, P.-A. 2008. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, 1096–1103.

Wang, X.; Zhang, R.; Shen, C.; Kong, T.; and Li, L. 2021. Dense contrastive learning for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3024–3033.

Wei, C.; Xie, L.; Ren, X.; Xia, Y.; Su, C.; Liu, J.; Tian, Q.; and Yuille, A. L. 2019. Iterative reorganization with weak spatial constraints: Solving arbitrary jigsaw puzzles for unsupervised representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1910–1919.

Wei, L.; Xiao, A.; Xie, L.; Chen, X.; Zhang, X.; and Tian, Q. 2020. Circumventing Outliers of AutoAugment with Knowledge Distillation. *arXiv preprint arXiv:2003.11342*.

Wu, Z.; Xiong, Y.; Yu, S. X.; and Lin, D. 2018. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3733–3742.

Xie, Z.; Lin, Y.; Zhang, Z.; Cao, Y.; Lin, S.; and Hu, H. 2021. Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16684–16693.

Zhang, H.; Cisse, M.; Dauphin, Y. N.; and Lopez-Paz, D. 2017. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*.

Zhang, R.; Isola, P.; and Efros, A. A. 2016. Colorful image colorization. In *European conference on computer vision*, 649–666. Springer.

Zhang, R.; Isola, P.; and Efros, A. A. 2017. Split-brain autoencoders: Unsupervised learning by cross-channel prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1058–1067.

Zhao, H.; Shi, J.; Qi, X.; Wang, X.; and Jia, J. 2017. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2881–2890.

Zhao, N.; Wu, Z.; Lau, R. W.; and Lin, S. 2020. What makes instance discrimination good for transfer learning? *arXiv preprint arXiv:2006.06606*.

Zhuang, C.; Zhai, A. L.; and Yamins, D. 2019. Local aggregation for unsupervised learning of visual embeddings. In *Proceedings of the IEEE International Conference on Computer Vision*, 6002–6012.