

Pixel-Wise Contrastive Distillation

Junqiang Huang^{†‡} Zichao Guo[†]
Shopee

jq.huang.work@gmail.com zichao.guo@shopee.com

Abstract

We present a simple but effective pixel-level self-supervised distillation framework friendly to dense prediction tasks. Our method, called Pixel-Wise Contrastive Distillation (PCD), distills knowledge by attracting the corresponding pixels from student’s and teacher’s output feature maps. PCD includes a novel design called SpatialAdaptor which “reshapes” a part of the teacher network while preserving the distribution of its output features. Our ablation experiments suggest that this reshaping behavior enables more informative pixel-to-pixel distillation. Moreover, we utilize a plug-in multi-head self-attention module that explicitly relates the pixels of student’s feature maps to enhance the effective receptive field, leading to a more competitive student. PCD **outperforms** previous self-supervised distillation methods on various dense prediction tasks. A backbone of ResNet-18-FPN distilled by PCD achieves 37.4 AP^{bbox} and 34.0 AP^{mask} on COCO dataset using the detector of Mask R-CNN. We hope our study will inspire future research on how to pre-train a small model friendly to dense prediction tasks in a self-supervised fashion.

1. Introduction

Self-supervised learning (SSL) has emerged as a promising pre-training method due to its remarkable progress on various computer vision tasks [24, 9, 21, 6, 23, 59]. Models pre-trained by SSL methods attain transfer performance akin to, or even surpassing that of their supervised pre-trained counterparts. However, this advancement of SSL appears to be confined to larger models. Small models, such as ResNet-18 [26], exhibit inferior linear probing accuracy as reported in [18, 60, 14]. Considering the necessity of small models for edge devices or resource constraint regime, it is much essential to tackle this problem.

Recently, the performance lag of small models has been effectively alleviated by *self-supervised distillation* [61, 39,

1, 18, 19, 60, 4, 14, 66, 54], where teachers’ (large pre-trained models) knowledge is transferred [5, 48, 2, 29] to students (small models) in a self-supervised learning fashion. Self-supervised distillation methods yield competitive performance for small models, especially on classification tasks (e.g., fine-grained and few-shot classification). Nevertheless, their improvement on dense prediction tasks like object detection and semantic segmentation is less pronounced than on classification tasks. This imbalance seemingly suggests that the favorable representations learned by teachers can only be *partially* transferred to students. A natural question arises: what obstructs students from inheriting the knowledge advantageous to dense prediction tasks? In this study, we seek the answers to this question from the following aspects.

First, the distillation signals of current self-supervised distillation methods are mostly at image-level, while the rich *pixel-level knowledge* is yet to be utilized. We argue that it is inefficient for small models to learn representations good for dense prediction tasks from image-level supervision¹. Driven by this, we here present a simple but effective pixel-level self-supervised distillation framework, **Pixel-Wise Contrastive Distillation** (PCD), which extends the idea of contrastive learning [22] by incorporating pixel-level knowledge distillation. PCD attracts the *corresponding* pixels from the student’s and the teacher’s output feature maps and separates the student’s pixels and the negative pixels of a memory queue [24]. With pixel-level distillation signal, PCD enables more efficient and adequate transfer of knowledge from the teacher to the student.

Second, it is not straightforward to distill pixel-level knowledge from the commonly adopted teachers pre-trained by image-level SSL methods [24, 9, 21, 6]. These teachers tend to project images into vectorized features, thus losing the spatial information which is *indispensable* to pixel-to-pixel distillation in our PCD. An intuitive practice to circumvent this issue would be to remove the well-trained projection/prediction head (a non-linear MLP attached to the backbone) and the global pooling layer. The

[†]Equal contribution

[‡]Corresponding author

¹On the other hand, large models pre-trained by image-level SSL methods like MoCo can be quite competitive on dense prediction tasks.

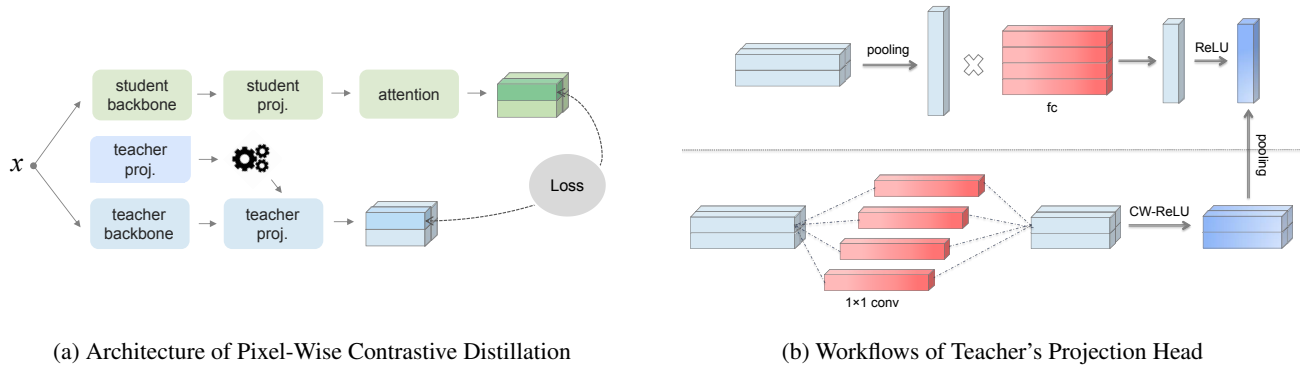



Figure 1: (a) is the specific architecture of Pixel-Wise Contrastive Distillation. Before distillation, the original teacher’s projection head is modified by SpatialAdaptor (represented by  in the figure). Distillation loss is the average contrastive loss computed over all corresponding pixel pairs of the student and the teacher. (b) depicts the workflows of teacher’s projection head before (top half) and after (bottom half) using SpatialAdaptor. The pooling layer on the far right is used for demonstrating the invariability of SpatialAdaptor. Best viewed in color.

distillation loss will be computed over the feature maps output by the backbone. However, we experimentally show the ineffectiveness of such a simplistic approach, implying that the projection/prediction head contains nonnegligible knowledge for pixel-level distillation. Towards the goal of leveraging this knowledge, we introduce a *SpatialAdaptor* to adapt the projection/prediction head used for encoding vectorized features to processing 2D feature maps while not changing the distribution of the output features. Though conceptually simple, the SpatialAdaptor is of great significance to pixel-level distillation.

Last, small models are innately weak in capturing information from the regions of large spans due to their smaller effective receptive fields (ERF) [37]. This natural deficiency prevents students from further imitating teachers at pixel-level. In this case, we append a multi-head self-attention (MHSA) module [49] to the student model, which *explicitly* relates the pixels within the student’s output feature maps. The addition of the MHSA module helps slightly enlarge the ERF of the small model, and consequently improves its transferring results. The MHSA will be deprecated in the fine-tuning phase. We think such gain without pain is very helpful for pre-training. We refer to Fig. 1a for a detailed depiction of PCD.

We comprehensively evaluate the effectiveness of PCD on several typical dense prediction tasks. PCD *surpasses* state-of-the-art results across all downstream tasks. Our results demonstrate the nontrivial advantages of PCD over competitive SSL methods designed for dense prediction tasks and previous image-level self-supervised distillation methods. Under the linear probing protocol, a ResNet-18 distilled by PCD attains 65.1% top-1 accuracy on ImageNet. These results highlight the superiority of pixel-level supervision signal for self-supervised distillation.

We also find that PCD is robust to the choices of pre-

trained teacher models and works well with various student backbone architectures. Students of larger backbones can compete with or even exceed the teacher on certain tasks, revealing an encouraging path for pre-training. These findings carry implications for future research and we hope our work inspires further investigation into self-supervised pre-training with small models.

2. Related Work

Pixel/region-level self-supervised learning aims to learn competitive representations specialized for dense prediction tasks. Following the philosophy of contrasting pixel/region-level features from different augmented views, these methods develop various rules to find the positive pairs.

Intuitive methods [40, 43, 56, 58, 62, 52] record the offsets and the scaling factors induced by geometric transformations (*e.g.*, cropping, resizing, and flipping) to locate the positive pairs of pixels/regions from different augmented views. In [27, 3], all pixels or regions within the original image are classified into some appropriate categories by a heuristic way or some unsupervised semantic segmentation methods. Any two pixels or regions from the same category form a positive pair. SoCo [51] and ORL [57] utilize the selective search [47] to identify numerous regions containing a single object and perform region-level contrastive learning based on these regions. DenseCL [50] and Self-EMD [35] pair the pixels of feature maps from different views according to some certain rules, *e.g.*, minimizing the cosine distances between pixels or finding the matching set with minimum earth mover’s distance [30].

Our PCD does not rely on sophisticated rules or preparations to pair pixels or regions. Instead, we directly contrast the feature maps output by the student and the teacher from the *same* view of an image, decoupling the

requirement for delicate augmentation policies from the design of pre-training framework.

Feature-based knowledge distillation transfers knowledge by matching the intermediate features of students and teachers, which are often not comparable due to the difference in shapes, *i.e.*, the number of channels and the spatial size. It is a common practice to reshape students’ features to have the same shape as teachers’ by a learnable module [44]. Some works [64, 33, 28, 8] transform both students’ and teachers’ features into tensors of the same shape. In PCD, shape alignment (especially with regard to the number of channels) is achieved by a non-linear projection head, which is a widely recognized technique in SSL for enhancing the quality of learned representations [9, 10, 21]. Additionally, in cases where the student and the teacher have feature maps of different spatial sizes, we employ a simple interpolation to complete the necessary alignment.

Self-supervised distillation transfers knowledge in a self-supervised learning fashion. CompRes [1] and SEED [18] propose to minimize the feature similarity distributions between students and teachers. DoGo [4] and DisCo [19] add a distillation branch for easing the optimization problem of small models during self-supervised pre-training. Previous works train students to classify images [61, 39] or minimize the intra-group distances [60] based on the clusterings generated by teachers. [14] simultaneously trains teachers and teacher from scratch. Students are guided by teachers’ on-the-fly clustering results. With these methods, the notorious problem that small models pre-trained by SSL methods face performance degradation has been partially solved. Our PCD is proposed to address the unsolved part—to improve the transferring results on dense prediction tasks.

3. Method

3.1. Pixel-Wise Contrastive Distillation

Unlike general knowledge distillation methods in supervised learning, self-supervised distillation does not involve with labeled data. The supervision only comes from teachers, yielding task-agnostic students that can be fine-tuned on various downstream tasks.

Image-level self-supervised distillation. Though varying dramatically in the specific rules of computing distillation loss, current self-supervised knowledge distillation methods [61, 39, 1, 18, 4, 19, 60, 14] share one thing in common—the supervision signals are all at image-level. Below, we describe a general formulation for these methods.

An input image is fed to the student’s and the teacher’s backbone, generating feature maps, $\mathbf{s} \in \mathbb{R}^{C_s \times H \times W}$ and

$\mathbf{t} \in \mathbb{R}^{C_t \times H \times W}$, respectively. C_s and C_t are the number of channels. H and W are the spatial sizes². These feature maps are then global average/max pooled into vectorized features. The distillation loss \mathcal{L} with respect to a single input image is defined by:

$$\mathcal{L} = \mathcal{L}(\phi(\mathbf{s}), \phi(\mathbf{t})), \quad (1)$$

where $\phi(\cdot)$ is the global pooling layer. Note that $\mathcal{L}(\cdot, \cdot)$ is a function composition³, but not a simple loss function like ℓ_2 distance. We let ϕ be global average pooling for the ease of analysis, *i.e.*, $\phi(\mathbf{x}) = \frac{1}{HW} \sum_i \mathbf{x}_i$. Here $i = (i_H, i_W)$ is a 2-tuple indexing the (i_H, i_W) -th pixel of feature maps.

Given this unified formulation, we consider the derivative with respect to the i -th pixel of student’s feature maps \mathbf{s}_i . By chain rule, we have:

$$\frac{\partial \mathcal{L}}{\partial \mathbf{s}_i} = \frac{\partial \mathcal{L}}{\partial \phi} \frac{\partial \phi}{\partial \mathbf{s}_i} = \frac{1}{HW} \frac{\partial \mathcal{L}}{\partial \phi}. \quad (2)$$

It is obvious to see that $\frac{\partial \mathcal{L}}{\partial \mathbf{s}_i}$ is a term *independent* from the position of the pixel. In other words, teacher’s guidance is not detailed at pixel-level. There probably exists huge disparity between student’s and teacher’s pixel-level features. Consequently, it is far beyond reach for students (small models) to inherit the competitive pixel-level knowledge possessed by teachers. We argue that this may be the very reason of students’ imbalanced performance on classification and dense prediction tasks.

Pixel-Wise Contrastive Distillation. Motivated by the above analysis, we propose a simple *pixel-level* self-supervised distillation framework, Pixel-Wise Contrastive Distillation (PCD). Our PCD transfers knowledge by attracting the positive pairs of pixels from students and teachers and repulsing the negative pairs.

Different from augmentation-invariant representation learning [46, 24, 38, 9, 58, 50], the positive pairs of PCD are from the *corresponding* pixels of student’s and teacher’s output feature maps for the same image, *i.e.*, $(\mathbf{s}_i, \mathbf{t}_i)$. Negative samples $\{\mathbf{n}^k | k = 1, \dots, K\}$ are stored in a queue in conformity with MoCo [24]. For an input image, we optimize the average contrastive loss of all output pixels:

$$\mathcal{L}(\mathbf{s}, \mathbf{t}) = \frac{1}{HW} \sum_i \ell(\mathbf{s}_i, \mathbf{t}_i, \{\mathbf{n}^k\}), \quad (3)$$

where ℓ stands for the contrastive loss function. We do not directly contrast \mathbf{s}_i and \mathbf{t}_i for they may have different dimensions (*i.e.*, the numbers of channels). Following

²It is reasonable to assume \mathbf{s} and \mathbf{t} have the same spatial size owing to the popular 32-stride design of convolutional network architectures. We also discuss about the situation where the student and the teacher produce features with different spatial sizedh.

³In CompRes [1], for example, \mathcal{L} is equivalent to first estimating the similarity distributions of the student and the teacher, then computing the KL divergence among them.

SimCLR [9], we append a projection head (a non-linear MLP φ parameterized by θ) to student’s backbone. Details of the projection head $\varphi(\cdot|\theta)$ will be given in Sec. 3.2. The projection head φ serves as aligning \mathbf{s}_i and \mathbf{t}_i in terms of dimension, and will be removed once training is accomplished. We denote the projected output $\varphi(\mathbf{s}_i|\theta)$ as \mathbf{s}_i^* for short. The concrete form of ℓ is:

$$\ell = -\log \frac{\exp(\mathbf{s}_i^{*\top} \mathbf{t}_i / \tau)}{\exp(\mathbf{s}_i^{*\top} \mathbf{t}_i / \tau) + \sum_k^K \exp(\mathbf{s}_i^{*\top} \mathbf{n}^k / \tau)}, \quad (4)$$

where τ is a temperature hyper-parameter. After back-propagation, teacher’s feature maps \mathbf{t} will be global pooled, ℓ_2 -normalized, and enqueued as a negative sample used for subsequent iterations. Here, we omit the ℓ_2 -normalization applied to \mathbf{s}_i^* and \mathbf{t}_i . So the inner product $\mathbf{s}_i^{*\top} \mathbf{t}_i$ equals to the cosine distance.

It is worth noting that PCD does not require \mathbf{s} and \mathbf{t} to have the same spatial size. In cases of \mathbf{s} and \mathbf{t} having mismatched spatial sizes, we perform bilinear interpolation on \mathbf{t} to match the spatial size of \mathbf{s} . Further discussions are in Sec. 4.3.

SpatialAdaptor. For the models pre-trained by image-level SSL methods (e.g., [24, 9, 21, 6]), their projection/prediction heads (henceforth projection head for simplicity) are the stacked fully-connected (fc) layers with batch normalization (BN) layers and ReLU in between. These projection heads only take the global pooled features as inputs. If adopting these models as teachers (a common practice in previous self-supervised distillation methods), one has to *remove* the global pooling layer and the projection heads to be compatible with our PCD. However, removing the well-trained projection heads will break the *integrity* of teachers. Such removal incurs knowledge loss, bringing in sub-optimal results for transfer learning. We will empirically verify this in Sec. 4.3.

To meet the demand of utilizing the fruitful knowledge of the projection heads, we propose a *SpatialAdaptor* to adapt the projection heads to processing 2D inputs. We next discuss a simple case where the projection head only contains a fc layer (f) and ReLU (σ), to demonstrate how the SpatialAdaptor works. Before using the SpatialAdaptor, the feature maps \mathbf{t} output by teacher’s backbone will be global average pooled and fed to the projection head:

$$\begin{aligned} \sigma(f(\phi(\mathbf{t}))) &= \sigma\left(f\left(\frac{1}{HW} \sum_i \mathbf{t}_i\right)\right) \\ &= \sigma\left(\frac{1}{HW} \sum_i f(\mathbf{t}_i)\right). \end{aligned} \quad (5)$$

The second equality holds being a consequence of f ’s linearity. By interchanging f and ϕ , f now acts on pixels rather than vectorized features. It follows that f can be *reformulated* into a 1×1 convolution (conv) layer with the stride of 1. This simple case does not consider the existence of any BN layer since fc and BN layer together can be fused into a linear function and represented by f .

Furthermore, we present the Channel-Wise ReLU (CW-ReLU) that masks out the channels whose mean are negative. Let σ^* denote CW-ReLU. By the definition, we have:

$$\sigma(\phi(\mathbf{x})) = \phi(\sigma^*(\mathbf{x})). \quad (6)$$

Eq. (6) means the global pooling layer and the activation function now are “*commutative*”. Combining Eq. (5) with this commutative property, we can interchange the global pooling layer and the activation function:

$$\sigma(f(\phi(\mathbf{t}))) = \phi(\sigma^*(f(\mathbf{t}))). \quad (7)$$

Note that ϕ in the right-hand side of Eq. (7) is only used for illustrating the *invariability* of the SpatialAdaptor. It will be omitted in actual use for maintaining spatial information.

Thus far, we have shown how a projection head composed of a fc layer and ReLU is adapted by the SpatialAdaptor to processing 2D feature maps while not changing the feature distribution (Fig. 1b). As such, the integrity of teachers is maintained, and our PCD is made compatible with the teachers pre-trained by various SSL methods. Though the case discussed above being so simple, the SpatialAdaptor can be easily generalized to the situation where the projection head is more complex (i.e., stacked with more fc layers and ReLU).

We are aware that the models pre-trained by pixel-level SSL methods [58, 50] have the projection heads processing 2D inputs. But this does *not* deprive the significance of the SpatialAdaptor, because we do not want too much constraints on teacher’s pre-training methods. Such accessibility can also be seen as a strength of our PCD.

Multi-head self-attention. Apart from the granularity of distillation signals, the *intrinsic properties* (e.g., limited capacity and smaller receptive field) of the students also play a role in self-supervised distillation. Consider the effective receptive field (ERF) [37] for an example. ERF measures how much each pixel contributes to the final prediction and has been proven to be closely related to the performance of abundant computer vision tasks [7, 13]. Intuitively, models with larger ERF are able to capture information from a bigger area of image, leading to more robust and reliable predictions.

According to [37], we draw the ERFs of ResNet-50 and ResNet-18 [26] in Fig. 2. We can see a clear contrast that ResNet-50 has larger ERF (larger bright region)

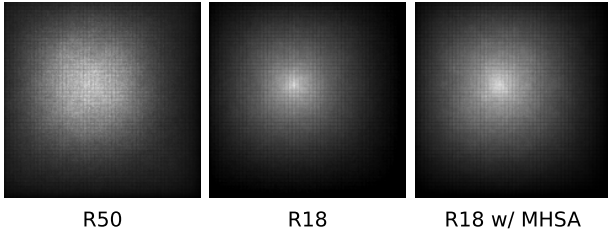


Figure 2: **Effective receptive field.** R50 and R18 are ResNet-50 and ResNet-18 respectively. R18 w/ MHSA stands for ResNet-18 enhanced by a multi-head self-attention module.

than ResNet-18. Therefore, it is *unrealistic* to expect small models to perfectly capture information from the regions of large spans like teachers do without outside help.

The definition of ERF indicates allowing more pixels to participate in predictions helps enlarge ERF. From this perspective, we can enhance the student model by *explicitly* relating all pixels right before making predictions. This is made possible by a multi-head self-attention (MHSA) module [49]. We introduce a MHSA module between student’s projection head and contrastive loss. Information from different pixels are aggregated together to make a more robust prediction. This module only induces a handful of memory and computation cost during pre-training, and does not influence the fine-tuning phase. Equipped with the MHSA module, the ERF of ResNet-18 is slightly enlarged (shown in the third picture of Fig. 2).

3.2. Baseline Settings

In this section, we provide the necessary details for implementing PCD.

Teacher. In most of our experiments, we adopt the ResNet-50 pre-trained by MoCo v3⁴ [12] for 1000 epochs as the teacher. We use the momentum updated branch (*i.e.*, `momentum_encoder` of the checkpoint) of MoCo v3 model. The projection head has the following structure: FC–BN–ReLU–FC–BN. The last BN has no affine parameters. We drop it because its static normalization statistics slow down the convergence. The global pooling layer of the backbone is removed. Modified by the SpatialAdaptor, the structure of the projection head becomes Conv–BN–CW–ReLU–Conv. The hidden and output dimension of the projection head is 4096 and 256 respectively. The shape of output feature maps is $256 \times 7 \times 7$ (channels \times spatial size). Other choices of teacher models are discussed in Sec. 4.3.

Student. By default, we use a ResNet-18 as the student’s

⁴The checkpoint can be found in <https://github.com/facebookresearch/moco-v3/blob/main/CONFIG.md>

backbone. It is followed by a projection head (φ mentioned in Sec. 3.1). We follow the asymmetric structure design in [21, 11, 12] to instantiate φ with two consecutive MLPs. The MLPs are structurally similar to the teacher’s projection head that is modified by the SpatialAdaptor, except that the activation function is ReLU. The hidden dimension of the student is equal to that of the teacher. We append a MHSA module [49] without position encoding [53] to the end of φ . The MHSA has 8 heads, each with an embedding dimension of 64. Each pixel of the input feature maps is regarded as a token. The attention values of all heads are concatenated and then projected by a 1×1 conv layer to match with the input features in dimension. The output of the MHSA module is used for computing the contrastive distillation loss. Other choices of student backbone architectures are discussed in Sec. 4.3.

Distilling. We perform self-supervised distillation on the ImageNet [45] training set. The image augmentation policy is the same as that proposed in [21], comprised of two distributions of augmentation. It generates a pair of augmented views for an input image. Our PCD computes distillation loss on each view and optimizes the mean of two losses. We observe this loss symmetrization brings about better convergence. We must point out that PCD with asymmetric loss also provides bright transfer performance (see in Sec. 4.2).

We use the LARS [63] optimizer to train for 100 epochs. The batch size is 1024. The base learning rate (lr) is set to be 1.0 and scaled by the linear scaling rule [20]: $lr \times \text{batch_size} / 256$. The learning rate is linearly increased to 4.0 for the first 10 epochs of training (warmup) and then decayed to 0 based on the cosine schedule. We use a weight decay of 0.00001 and a momentum of 0.9. All biases and the affine parameters of BN layers are excluded from the weight decay. We set the temperature of contrastive loss to be 0.2. The queue storing negative samples has a capacity of 65536.

4. Experiments

4.1. Evaluation Setup

Here, we provide some background for our fine-tuning experiments. We evaluate our PCD on VOC [17] object detection, COCO [34] object detection and instance segmentation, and CityScapes [15] semantic segmentation. The codebase for evaluation is Detectron2 [55]. We strictly follow the fine-tuning settings proposed in [24, 58] for fair comparisons. We also provide the linear probing accuracy on ImageNet.

It has been a notorious problem that fine-tuning LARS-trained models with optimization hyper-parameters best selected for SGD-trained counterparts yields sub-optimal performance [11, 32]. To address this, recent work [32]

	ImageNet	VOC 07		VOC 07+12		COCO-C4		COCO-FPN		CityScapes
	Acc	AP ₅₀	AP	AP ₅₀	AP	AP ^{bbbox}	AP ^{mask}	AP ^{bbbox}	AP ^{mask}	IoU
Teacher	74.6	77.8	48.1	83.0	56.7	37.4	32.8	40.7	36.9	73.5
ImageNet Supervised	69.8	70.0	38.2	76.9	47.3	30.7	28.0	36.3	33.0	70.2
MoCo v2 [10]	48.7	69.7	40.2	77.5	49.7	30.7	27.9	35.0	31.8	70.4
PixPro [58]	41.4	71.5	42.3	78.5	51.1	30.9	28.1	35.8	32.6	70.3
CompRes [1]	63.9	71.3	41.2	78.4	50.4	31.4	28.4	35.7	32.4	70.3
DisCo [19]	63.5	62.5	30.7	72.6	40.1	28.2	25.8	36.0	32.8	69.3
BINGO [60]	64.2	70.4	39.9	77.8	49.3	31.1	28.2	36.2	32.8	71.0
PCD, asymmetric	64.2	72.7	42.7	79.3	51.5	31.9	28.8	37.0	33.7	71.6
PCD	65.1	73.0	43.2	79.4	52.1	32.2	29.0	37.4	34.0	71.8

Table 1: **Comparing different pre-trained models.** All pre-trained models adopt ResNet-18 as backbone. ImageNet supervised pre-trained model is from the model zoo of PyTorch [41]. Other pre-trained models are from our reproductions built on their officially released codes. For fair comparisons, we pre-trained these models for 100 epochs. We use the ResNet-50 pre-trained by MoCo v3 as teacher for all self-supervised distillation methods. The best results are marked in **bold**, and the second best are marked in *gray* (exclusive of the teacher).

proposes NormRescale to scale the norm of LARS-trained weights by a specific anchor (*e.g.*, the norm of SGD-trained weights or a constant number). It helps the LARS-trained models fit to the optimization strategy of fine-tuning. When fine-tuning C4 or FCN backbones pre-trained by PCD, we employ this technique to multiply the weights by a constant 0.25. Multiplying a constant is an efficient choice for not introducing extra training cost.

VOC object detection. We use a C4 backbone with Faster R-CNN [42] detector. We evaluate the pre-trained models under two fine-tuning settings. The first is to train on `trainval2007` set for 9k iterations, and the second is to train on `trainval07+12` set for 24k iterations. We use the same fine-tuning settings as per [24]. Fine-tuned models are evaluated on `test2007` set. For better reproducibility, we report the average AP₅₀ and AP over 5 runs.

COCO object detection and instance segmentation. We use two types of backbone, C4 and FPN, for fine-tuning on COCO dataset. The detector is Mask R-CNN [25]. Pre-trained models are fine-tuned on `train2017` set according to the $1\times$ optimization setting (about 12 COCO epochs). We report AP^{bbbox} for object detection and AP^{mask} for instance segmentation on `val2017` set.

CityScapes semantic segmentation. We implement a FCN-like [36] structure based on pre-trained backbones. A newly initialized BN layer is added to the end of pre-trained backbones for helping optimization. Subsequently, we append two atrous convolutional blocks, each with a 3×3 conv layer of 256 output channels, a BN layer, and ReLU. The conv layers have stride 1, dilation 6, and padding 6. The prediction layer is a 1×1 conv layer with 19 output

channels (19 classes), whose outputs are then bilinearly interpolated to match the size of input images. Fine-tuning takes 90k iterations on `train_fine` set. More detailed information of fine-tuning can be found in Appendix. We report the average IoU on `val` set over 5 runs.

Linear probing in ImageNet. We freeze the backbones pre-trained by PCD and train a linear classifier on the ImageNet training set. We use nesterov SGD to train for 100 epochs. The batch size is 1024, and the learning rate is 0.8 (base $lr = 0.2$). The learning rate will decay to 0 according to the cosine schedule without restart. The momentum is 0.9, and the weight decay is 0. We use a vanilla image augmentation policy containing random cropping, resizing to 224×224 , and horizontal flipping. We report the single-crop classification accuracy on the ImageNet validation set.

4.2. Main Results

In Tab. 1, we present the fine-tuning results of the following pre-training methods: supervised pre-training, SSL methods, previous competitive self-supervised distillation methods, and our PCD. We notice that PixPro [58] (a SSL method designed for dense prediction tasks) outperforms the self-supervised distillation methods on most tasks. This observation confirms our hypothesis that small models are *difficult* to learn pixel-level knowledge from image-level pretext tasks, even with distillation signals, further justifying the necessity of our method.

Our PCD shows impressive generalization capacity: it surpasses all competitors on each dense prediction task. We achieves 37.4 AP^{bbbox} and 34.0 AP^{mask} on COCO using the Mask R-CNN detector and the ResNet-18-FPN backbone,

	VOC 07+12		COCO-FPN		CityScapes
	AP ₅₀	AP	AP ^{bbox}	AP ^{mask}	IoU
image-level	78.7	49.8	36.5	33.3	71.1
pixel-level	79.4	52.1	37.4	34.0	71.8

Table 2: **Pixel-level vs. image-level.** We compare PCD to image-level contrastive distillation. The backbone is ResNet-18 and pre-trained for 100 epochs. The best results are marked in **bold**.

	VOC 07+12		COCO-FPN		CityScapes
	AP ₅₀	AP	AP ^{bbox}	AP ^{mask}	IoU
(a)	76.0	48.8	36.5	33.2	70.1
(b)	77.9	50.6	36.8	33.6	70.7
(c)	77.3	50.1	36.3	33.1	69.6
(d)	77.8	50.7	36.6	33.2	70.1
ours	79.4	52.1	37.4	34.0	71.8

Table 3: **Ablations on SpatialAdaptor.** We compare four variants of PCD (a-d) to examine the necessity of SpatialAdaptor. The backbone is ResNet-18 and pre-trained for 100 epochs. The best results are marked in **bold**.

emerging as the first pre-training method exceeding the supervised pre-trained model on this benchmark. Under the linear probing protocol, our PCD also achieves decent top-1 accuracy (65.1%), which makes PCD a well-rounded self-supervised distillation method.

We notice that some competitors (*e.g.*, supervised learning, MoCo v2 [10], and CompRes [1]) are pre-trained by asymmetric loss. Here, we provide an asymmetric variant of PCD to exclude the effect of loss symmetrization. The change is simple: we adopt the symmetric augmentation policy as per [21], and sample one augmented view from each input image during training. Indeed, the symmetric loss endows PCD with better performance, but asymmetric variant still achieves the second best results on all tasks (marked as gray in Tab. 1). The *nontrivial* advantages of PCD against other self-supervised distillation methods have confirmed the fact that pixel-level distillation signals are the key to transferring knowledge conducive to dense prediction tasks.

4.3. Ablation Experiments

We perform extensive ablation experiments to analyze PCD. Unless specified, we adopt the training settings mentioned in Sec. 3.2.

Pixel-level vs. image-level. To directly compare pixel-level and image-level distillation, we develop an image-level variant of PCD. Based on PCD, this variant vectorizes student’s and teacher’s output feature maps by an extra

	VOC 07+12		COCO-FPN		CityScapes
	AP ₅₀	AP	AP ^{bbox}	AP ^{mask}	IoU
w/o MHSA	79.2	51.8	37.1	33.7	71.6
extra pred.	79.1	52.0	37.2	33.7	71.2
ours	79.4	52.1	37.4	34.0	71.8

Table 4: **Ablations on MHSA.** We ablate the MHSA module. “extra pred.” stands for replacing the MHSA module by an extra prediction head. The backbone is ResNet-18 and pre-trained for 100 epochs. The best results are marked in **bold**.

global average pooling layer and computes contrastive loss on these vectorized features. It has highly competitive results (Tab. 2) like those image-level self-supervised distillation methods in Tab. 1, revealing the effectiveness of contrastive loss used for self-supervised distillation. But it still *lags behind* the original PCD on all downstream tasks. This gap further confirms the importance of pixel-level distillation signal.

Ablation on SpatialAdaptor. We examine the necessity of the SpatialAdaptor for learning competitive representations. Without resorting to the SpatialAdaptor, we remove teacher’s projection head (along with the global pooling layer) and simply use the feature maps output by teacher’s backbone to compute contrastive loss (variant (a) in Tab. 3). The evaluation metrics AP₅₀ and AP on VOC are rather low. This variant overlooks the fact that teacher’s feature maps have numerous zeros (the characteristic of ReLU) while student’s feature maps do not. Contrasting two features from different distributions naturally leads to sub-optimal results.

For more reasonable comparisons, we introduce two more variants extended from variant (a). Variant (b) adds ReLU after the MHSA module. Variant (c) removes the MHSA module and adds ReLU after student’s projection head. Overall, these two variants (Tab. 3 (b-c)) are still significantly worse than PCD. And they are no better than the image-level self-supervised distillation methods in Tab. 1 and Tab. 2. We argue that preserving the integrity of teachers (with the help of the SpatialAdaptor) is of vital importance to pixel-level distillation. Otherwise, it will notably lower the quality of learned representations.

Additionally, we study the impact of invariability of SpatialAdaptor by replacing CW-ReLU with ReLU (variant (d)). We observe that keeping the distribution of teacher’s learned features unchanged has massive gains on dense prediction tasks (Tab. 3 (d)). In sum, the SpatialAdaptor is an essential component of PCD, enabling more *informative* pixel-wise distillation from teachers pre-trained by image-level SSL methods.

backbone	pre-training method	VOC 07		VOC 07+12		COCO-C4		COCO-FPN		CityScapes
		AP ₅₀	AP	AP ₅₀	AP	AP ^{bbox}	AP ^{mask}	AP ^{bbox}	AP ^{mask}	IoU
ResNet-50	SwAV	72.1	41.7	79.0	51.2	31.9	28.8	37.1	33.7	70.7
ResNet-50	BYOL	71.7	42.6	78.9	51.5	32.0	29.0	37.0	33.7	71.0
ResNet-50	Barlow Twins	70.7	41.6	78.4	50.7	31.3	28.4	35.9	32.7	70.9
ViT-Base	MoCo v3	70.6	41.7	78.1	51.1	31.3	28.5	36.6	33.6	71.4

Table 5: **Fine-tuning results of PCD with different teachers.** The student backbone is ResNet-18. All teacher models are from the officially released checkpoints. We refer to Appendix for more details about these teachers.

	pre-training method	VOC 07		VOC 07+12		COCO-C4		COCO-FPN		CityScapes
		AP ₅₀	AP	AP ₅₀	AP	AP ^{bbox}	AP ^{mask}	AP ^{bbox}	AP ^{mask}	IoU
ResNet-34	supervised	74.8	45.0	81.0	55.0	37.7	32.9	39.4	35.6	71.9
	PCD	76.3	49.1	82.0	58.2	38.2	33.7	40.9	37.0	73.2
ResNet-50	supervised	75.2	44.5	81.5	54.1	38.2	33.5	40.2	36.3	72.3
	PCD	77.0	49.0	82.8	57.7	40.1	34.9	42.4	38.1	73.3
MobileNet v3	supervised	67.9	36.7	76.2	46.4	30.6	27.9	35.8	32.7	68.3
	PCD	73.9	40.8	79.3	49.9	32.8	29.6	37.7	34.3	69.0

Table 6: **Fine-tuning results of PCD with different student backbones.** The teacher used for PCD is ResNet-50 pre-trained by MoCo v3. The student backbones are ResNet-18, ResNet-34, ResNet-50, and MobileNet v3 (Large). We also fine-tune the supervised pre-trained counterparts for contrast. We refer to Appendix for more implementation details. The best results for each backbone are marked as **bold**.

Ablation on multi-head self-attention. We ablate the MHSA module in Tab. 4. PCD without the MHSA module meets slight performance drop. A plausible explanation for the positive impact of the MHSA module is that it works like a prediction head to promote the quality of learned representations [21, 11, 12]. We thus supersede the MHSA module by an extra prediction head of roughly the same number of parameters. This substitution does not bring any improvement (Tab. 4), suggesting that explicitly relating pixels is useful for PCD. The MHSA module only adds a small computational overhead to the pre-training phase, but it consistently benefits to various downstream tasks. We therefore regard it as a necessary part to PCD.

Different teachers. Both MoCo v3 and PCD are trained with contrastive loss. To *exclude* the positive or negative effect induced by optimizing the same type of loss, we consider using teachers pre-trained by SwAV [6], BYOL [21], and Barlow Twins [65]. The fine-tuning results in Tab. 5 show that these teachers can also inspire favorable representations. The effectiveness of PCD is not strictly correlated to the teacher model pre-trained by MoCo v3. It can be concluded that PCD is *robust* to the choices of teacher models.

Beyond typical convolutional architectures, we try using ViT [16] as the teacher to study the effect of cross-architectures distillation. There is an innate obstacle for ResNet-18 to imitate ViT at pixel-level, since they differ

in the resolutions⁵ of output feature maps. Therefore, we downsample the output of ViT by a 2×2 average pooling layer with a stride of 2. Another solution would be to employ strided convolution in the projection head of student model. We leave it to be a topic of future research. It leads to acceptable results on dense prediction tasks, whereas unsatisfying linear top-1 accuracy (57.6%) on ImageNet. We believe cross-architectures distillation (between CNNs and transformers) is a noteworthy problem for future research.

Different students. We investigate the effectiveness of PCD on different student backbones: ResNet-34, ResNet-50, and MobileNet v3 (Large) [31]. Compared to supervised pre-training, our PCD consistently outperforms on all backbones (Tab. 6). A clear *trend* is that backbones with larger capacity (from ResNet-18 to ResNet-50) have better transfer performance. A distilled ResNet-34 or ResNet-50 can rival or even beat the teacher (referred to Tab. 1) on some downstream tasks, marking the practicability of our PCD.

5. Conclusion

In this paper, we study the notorious problem that small models pre-trained by SSL methods faces performance

⁵ViT-Base has 14×14 output patches for a 224×224 input, and we treat each patch as a pixel.

degradation on downstream tasks, especially on dense prediction tasks. We find it difficult for small models to learn pixel-level knowledge from image-level pretext tasks, even with distillation signals. To address this problem, we propose a simple but effective self-supervised distillation framework friendly to dense prediction tasks. Given the remarkable performance of PCD, we believe it points out a practical solution to pre-training small models in a self-supervised fashion.

6. Acknowledgements

Junqiang Huang would like to thank his mother Caixia Xu and his wife Yuwei Lin for their support and help along the way. We are grateful for the selfless assistance offered by our friend, Chengpeng Chen.

References

- [1] Soroush Abbasi Koohpayegani, Ajinkya Tejankar, and Hamed Pirsiavash. Compress: Self-supervised learning by compressing representations. *Advances in Neural Information Processing Systems*, 33:12980–12992, 2020. 1, 3, 6, 7
- [2] Jimmy Ba and Rich Caruana. Do deep nets really need to be deep? *Advances in neural information processing systems*, 27, 2014. 1
- [3] Yutong Bai, Xinlei Chen, Alexander Kirillov, Alan Yuille, and Alexander C Berg. Point-level region contrast for object detection pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16061–16070, 2022. 2
- [4] Prashant Bhat, Elahe Arani, and Bahram Zonooz. Distill on the go: Online knowledge distillation in self-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2678–2687, 2021. 1, 3
- [5] Cristian Bucilua, Rich Caruana, and Alexandru Niculescu-Mizil. Model compression. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 535–541, 2006. 1
- [6] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems*, 33:9912–9924, 2020. 1, 4, 8
- [7] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 4
- [8] Pengguang Chen, Shu Liu, Hengshuang Zhao, and Jiaya Jia. Distilling knowledge via knowledge review. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5008–5017, 2021. 3
- [9] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 1, 3, 4
- [10] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 3, 6, 7
- [11] 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, 8
- [12] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9640–9649, 2021. 5, 8
- [13] Yunpeng Chen, Haoqi Fan, Bing Xu, Zhicheng Yan, Yan-nis Kalantidis, Marcus Rohrbach, Shuicheng Yan, and Jiashi Feng. Drop an octave: Reducing spatial redundancy in convolutional neural networks with octave convolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3435–3444, 2019. 4
- [14] Hee Min Choi, Hyoa Kang, and Dokwan Oh. Unsupervised representation transfer for small networks: I believe i can distill on-the-fly. *Advances in Neural Information Processing Systems*, 34:24645–24658, 2021. 1, 3
- [15] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 5
- [16] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 8
- [17] Mark Everingham, SM Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111(1):98–136, 2015. 5
- [18] Zhiyuan Fang, Jianfeng Wang, Lijuan Wang, Lei Zhang, Yezhou Yang, and Zicheng Liu. Seed: Self-supervised distillation for visual representation. *arXiv preprint arXiv:2101.04731*, 2021. 1, 3
- [19] Yuting Gao, Jia-Xin Zhuang, Ke Li, Hao Cheng, Xi-aowei Guo, Feiyue Huang, Rongrong Ji, and Xing Sun. Disco: Remedy self-supervised learning on lightweight models with distilled contrastive learning. *arXiv preprint arXiv:2104.09124*, 2021. 1, 3, 6
- [20] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large mini-batch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017. 5
- [21] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent—a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020. 1, 3, 4, 5, 7, 8

- [22] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE, 2006. 1
- [23] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022. 1
- [24] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 1, 3, 4, 5, 6
- [25] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 6
- [26] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1, 4
- [27] Olivier J Hénaff, Skanda Koppula, Jean-Baptiste Alayrac, Aaron Van den Oord, Oriol Vinyals, and João Carreira. Efficient visual pretraining with contrastive detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10086–10096, 2021. 2
- [28] Byeongho Heo, Jeesoo Kim, Sangdoon Yun, Hoyjin Park, Nojun Kwak, and Jin Young Choi. A comprehensive overhaul of feature distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1921–1930, 2019. 3
- [29] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7), 2015. 1
- [30] Frank L Hitchcock. The distribution of a product from several sources to numerous localities. *Journal of mathematics and physics*, 20(1-4):224–230, 1941. 2
- [31] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1314–1324, 2019. 8
- [32] Junqiang Huang, Xiangwen Kong, and Xiangyu Zhang. Revisiting the critical factors of augmentation-invariant representation learning. In *European Conference on Computer Vision*, pages 42–58. Springer, 2022. 5
- [33] Jangho Kim, SeongUk Park, and Nojun Kwak. Paraphrasing complex network: Network compression via factor transfer. *Advances in neural information processing systems*, 31, 2018. 3
- [34] 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 *European conference on computer vision*, pages 740–755. Springer, 2014. 5
- [35] Songtao Liu, Zeming Li, and Jian Sun. Self-emd: Self-supervised object detection without imagenet. *arXiv preprint arXiv:2011.13677*, 2020. 2
- [36] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 6
- [37] Wenjie Luo, Yujia Li, Raquel Urtasun, and Richard Zemel. Understanding the effective receptive field in deep convolutional neural networks. *Advances in neural information processing systems*, 29, 2016. 2, 4
- [38] 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. 3
- [39] Mehdi Noroozi, Ananth Vinjimoor, Paolo Favaro, and Hamed Pirsiavash. Boosting self-supervised learning via knowledge transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9359–9367, 2018. 1, 3
- [40] Pedro O O Pinheiro, Amjad Almahairi, Ryan Benmalek, Florian Golemo, and Aaron C Courville. Unsupervised learning of dense visual representations. *Advances in Neural Information Processing Systems*, 33:4489–4500, 2020. 2
- [41] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. 6
- [42] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 6
- [43] 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
- [44] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014. 3
- [45] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015. 5
- [46] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *European conference on computer vision*, pages 776–794. Springer, 2020. 3
- [47] Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders. Selective search for ob-

- ject recognition. *International journal of computer vision*, 104(2):154–171, 2013. [2](#)
- [48] Ruth Urner, Shai Shalev-Shwartz, and Shai Ben-David. Access to unlabeled data can speed up prediction time. In *ICML*, 2011. [1](#)
- [49] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. [2](#), [5](#)
- [50] Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Dense contrastive learning for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3024–3033, 2021. [2](#), [3](#), [4](#)
- [51] Fangyun Wei, Yue Gao, Zhirong Wu, Han Hu, and Stephen Lin. Aligning pretraining for detection via object-level contrastive learning. *Advances in Neural Information Processing Systems*, 34:22682–22694, 2021. [2](#)
- [52] Xin Wen, Bingchen Zhao, Anlin Zheng, Xiangyu Zhang, and Xiaojuan Qi. Self-supervised visual representation learning with semantic grouping. *arXiv preprint arXiv:2205.15288*, 2022. [2](#)
- [53] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. Cvt: Introducing convolutions to vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22–31, 2021. [5](#)
- [54] Kan Wu, Jinnian Zhang, Houwen Peng, Mengchen Liu, Bin Xiao, Jianlong Fu, and Lu Yuan. Tinyvit: Fast pretraining distillation for small vision transformers. *arXiv preprint arXiv:2207.10666*, 2022. [1](#)
- [55] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. <https://github.com/facebookresearch/detectron2>, 2019. [5](#)
- [56] Tete Xiao, Colorado J Reed, Xiaolong Wang, Kurt Keutzer, and Trevor Darrell. Region similarity representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10539–10548, 2021. [2](#)
- [57] Jiahao Xie, Xiaohang Zhan, Ziwei Liu, Yew Soon Ong, and Chen Change Loy. Unsupervised object-level representation learning from scene images. *Advances in Neural Information Processing Systems*, 34:28864–28876, 2021. [2](#)
- [58] Zhenda Xie, Yutong Lin, Zheng Zhang, Yue Cao, Stephen Lin, and Han Hu. Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16684–16693, 2021. [2](#), [3](#), [4](#), [5](#), [6](#)
- [59] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9653–9663, 2022. [1](#)
- [60] Haohang Xu, Jiemin Fang, Xiaopeng Zhang, Lingxi Xie, Xinggang Wang, Wenrui Dai, Hongkai Xiong, and Qi Tian. Bag of instances aggregation boosts self-supervised distillation. In *International Conference on Learning Representations*, 2021. [1](#), [3](#), [6](#)
- [61] Xueting Yan, Ishan Misra, Abhinav Gupta, Deepti Ghadiyaram, and Dhruv Mahajan. Clusterfit: Improving generalization of visual representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6509–6518, 2020. [1](#), [3](#)
- [62] Ceyuan Yang, Zhirong Wu, Bolei Zhou, and Stephen Lin. Instance localization for self-supervised detection pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3987–3996, 2021. [2](#)
- [63] Yang You, Igor Gitman, and Boris Ginsburg. Large batch training of convolutional networks. *arXiv preprint arXiv:1708.03888*, 2017. [5](#)
- [64] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv preprint arXiv:1612.03928*, 2016. [3](#)
- [65] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pages 12310–12320. PMLR, 2021. [8](#)
- [66] Kai Zheng, Yuanjiang Wang, and Ye Yuan. Boosting contrastive learning with relation knowledge distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 3508–3516, 2022. [1](#)