Self-Supervised Representation Learning by Rotation Feature Decoupling

Zeyu Feng Chang Xu Dacheng Tao
UBTECH Sydney AI Centre, School of Computer Science, FEIT,
University of Sydney, Darlington, NSW 2008, Australia
zfen2406@uni.sydney.edu.au, {c.xu, dacheng.tao}@sydney.edu.au

Abstract

We introduce a self-supervised learning method that focuses on beneficial properties of representation and their abilities in generalizing to real-world tasks. The method incorporates rotation invariance into the feature learning framework, one of many good and well-studied properties of visual representation, which is rarely appreciated or exploited by previous deep convolutional neural network based self-supervised representation learning methods. Specifically, our model learns a split representation that contains both rotation related and unrelated parts. We train neural networks by jointly predicting image rotations and discriminating individual instances. In particular, our model decouples the rotation discrimination from instance discrimination, which allows us to improve the rotation prediction by mitigating the influence of rotation label noise, as well as discriminate instances without regard to image rotations. The resulting feature has a better generalization ability for more various tasks. Experimental results show that our model outperforms current state-of-the-art methods on standard self-supervised feature learning benchmarks.

1. Introduction

Deep neural networks, especially convolutional neural networks (ConvNets), have led to breakthroughs in the field of computer vision. Given large scale manually labeled image datasets, *e.g*. ImageNet, ConvNets can be well trained by back propagation and achieve state-of-the-art performance on many tasks such as image classification [25, 45] and object detection [31]. Rich representations extracted by these networks often serve as good general-purpose features not only for the task where the network was trained, but also for many other vision tasks like semantic segmentation [33] and visual question answering [2]. However, training deep neural networks in a fully supervised manner requires a tremendous amount of efforts on manual labeling, which could be infeasible in some real-world scenarios.

As an alternative to supervised feature learning, unsupervised methods that do not rely on expensive and time-consuming human labeling are receiving increasingly more attention. The recently emerged self-supervised learning paradigm [10, 43, 52, 27, 37] is a scalable and promising solution for learning useful general-purpose visual representations. These methods used to exploit the structural information of data itself and define pretext tasks that relates to the final application of the learned features to train neural network. In pretext tasks, supervisory signals can be easily developed without significant human efforts, and thus massive readily available images can be applied for training.

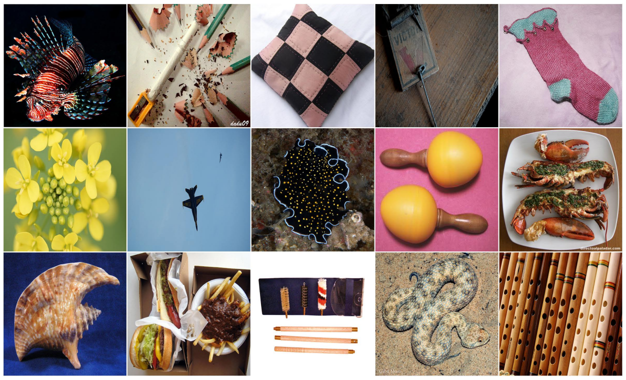


Figure 1: Examples of some rotation agnostic images in ImageNet. The default orientations of these images are ambiguous.

Over the last few years, many different pretext tasks have been proposed for self-supervised learning. For example, one category of these methods tries to recover one part of the data from the other part itself [43, 28, 53]. While a shortcoming of these approach is the requirement of reconstructing and predicting image pixel values, which often need significant computational resources. Deep neural networks can also be trained to differentiate original images from restored incomplete images [21]. However, generating synthetic images is not always an easy feat. Siamese network architecture has been investigated in self-supervised learning [2, 36, 53], but memory consumption is usually huge. Another different but popularly adopted strategy is to discover supervisory signals in videos like tracking image-patches [47] and sorting frame sequences [30].

Most of existing works focus on designing various pretext tasks, while seldom caring about what properties are owned by the learned representations and whether they are indeed beneficial for the generalization in real-world tasks. For example, high-level representations should convey a clear explanation or certain dependencies of factors of variation [5]. A recent attempt is to predict image rotations [17]. Features learned in this method can generalize well in various tasks, and achieve state-of-the-art performance. However, these features are discriminative with rotation transformation and therefore cannot benefit vision tasks that are in favor of rotation invariance. Moreover, it is instructive to note that not all examples are rotation determinable in practice. The orientation of an image is ambiguous not only for round objects, but also for many other objects in images that are orientation agnostic, for instance, some objects viewed from top or in symmetrical shape, as shown in Figure 1. Rotating these objects would not significantly influence our description or understanding.

In this paper, we present a new self-supervised learning algorithm that decouples representations through a rotation prediction task and an instance discrimination task. The learned example feature consists of two ingredients that are rotation discriminative and rotation unrelated, respectively. Rotation discriminative features can be discovered by predicting image rotations, which is simple yet effective and achieves state-of-the-art results on some benchmarks [17]. Regarding those orientation agnostic images in the dataset, automatically assigned rotation labels usually contain noise, which naturally leads to a positive unlabeled learning problem. Original images in the default orientation are positive instances while the rotated copies are unlabeled instances, which can be positive or negative. If the transformation of a rotated copy cannot be recognized unequivocally, we treat it as a positive instance with default orientation in the unlabeled set (See Figure 1 in supplementary material). On the other hand, we learn rotation unrelated features by penalizing the distance difference between features of the same image under different rotations. Non-parametric method is applied to distinguish different instances based on these rotation unrelated features. Hence, the features would have the discriminative ability on instance level.

To demonstrate the effectiveness of our self-supervised learning method, we conduct experiments on standard feature transfer learning benchmarks. We perform ablation studies to examine individual components in our model and different configurations. We also test the features on rotated dataset. Experimental results suggest that it is necessary to investigate rotation related and unrelated features. Features learned in our method outperform those of the state-of-the-art methods on many tasks including linear classification on ImageNet and Places, as well as classification, detection and segmentation on PASCAL VOC.

2. Related work

This work relates to several topics in machine learning and computer vision: self-supervised learning, positive unlabeled (PU) learning and image rotation invariance.

Self-supervised learning. Self-supervised learning constructs some supervisory signals directly computed from the input data. For example, some methods try to recover part of the data itself, such as image completion [43], image colorization [52, 27, 28] and channel prediction [53]. Others leverage concept information in images and then construct constraints, such as image patch position [10, 36], solving jigsaw puzzle [37], counting [38], rotation [17] and instance discrimination [13, 48]. Methods relying on adversarial training include [12] and [21]. Noroozi *et al.* [39] and Caron *et al.* [6] use clustering approach to generate pseudo-labels. Apart from single task, Doersch and Zisserman [11] and Ren and Lee [44] also consider using several tasks together to obtain performance boost. For videos, some examples of supervisory signals are: egomotion [1, 42], temporal coherence [47, 30] and sound [41]. Our method is based on predicting image rotations [17] and considers properties owned by the learned representations. We focus more on rotation related and unrelated property.

Positive unlabeled learning. Unlabeled data in PU learning are generally treated as negative examples, which means only the observed negative examples contain noisy labels [14]. Many methods look into the relationship between the conditional probability and its estimation to model the mislabeled rate [46, 40]. Then the mislabeled rate can be used to handle noisy observed negative examples by various ways such as excluding examples with low confidence [40], labeling examples with high confidence [49, 23, 19] or reweighting examples [14, 35, 32]. However, PU learning methods that possess good theoretical properties may not properly scale up to deep networks trained with millions of examples. In this work, we formulate the task of predicting image rotations as a PU learning problem and deal with the label noise by applying weights to unlabeled examples.

Rotation invariance. Many classical hand-crafted features like SIFT [34] and RIFT [29] for computer vision are insensitive to certain rotation transformations. For recent ConvNets based feature learning, some carefully designed network structure, such as G-CNNs [7] and Warped Convolutions [20], exhibit excellent results in learning rotational invariant features. Invariance to arbitrary set of transformations can be realized through data augmentation. Laptev *et al.* [26] extract max-pooled activation of multiple rotational copies of images. Dieleman *et al.* [9] expand the feature maps by combining various transformed features. These invariant representation learning methods are mainly trained in supervised tasks. We aim to learn compound features that contain rotation unrelated part in an unsupervised way. Our

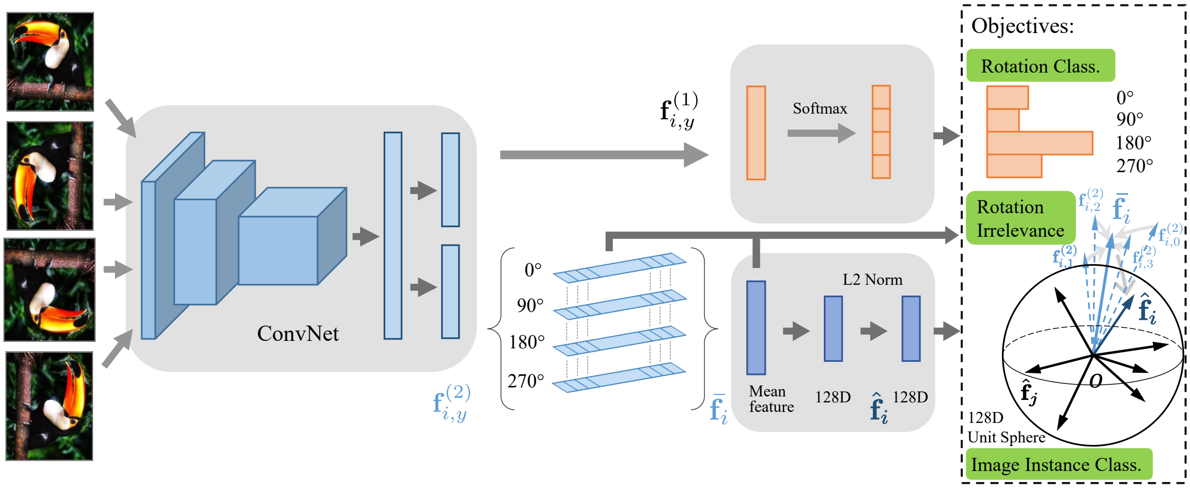


Figure 2: Illustration of the proposed method. The neural network outputs a decoupled semantic feature containing rotation related and unrelated parts. The first part is trained by predicting image rotations. Noises in rotation labels are modeled as a PU learning problem, which learns instance weights to reduce the influence of rotation ambiguous images. The other part is trained with a distance penalty loss to enforce rotation irrelevance together with an instance discrimination task by using non-parametric classification.

method also relies on multiple rotational copies of the data, while we utilize this rotation information efficiently for two decoupled unsupervised tasks.

3. Rotation feature decoupling

In this section, we first review the method of predicting image rotations (RotNet) [17], and then reformulate it as a positive-unlabeled learning problem that mitigates the innate defects in the design of this pretext task. We describe in detail of our rotation feature decoupling approach and give our complete model (See Figure 2).

3.1. Image rotation prediction

ConvNets are particularly powerful in mapping raw image to a semantically meaningful feature vector, but they are often trained using images and their corresponding ground truth labels. To obtain a general-purpose feature for an image in an unsupervised way, RotNet investigates geometric transformations of image, specifically rotations of image by multiples of 90 degrees, as supervisory signals, and trains ConvNet to predict their transformation [17]. Semantically meaningful representation can therefore be encoded in the feature maps of higher layers of the ConvNet.

Given a training dataset $S = \{X_i\}_{i=1}^{N}$ of N images, the RotNet defines a set of rotational transformation $G = \{g(X;y)\}_{y=1}^{K}$ for each image X. We denote the i-th image with the y-th rotation by $X_{i,y}$, where $X_{i,y} = g(X_i;y)$. A ConvNet model $F(\cdot;\boldsymbol{\theta})$ is trained to classify each rotated image to one of the transformations. The objective is:

$$\min_{\boldsymbol{\theta}} \frac{1}{NK} \sum_{i=1}^{N} \sum_{y=1}^{K} l(F(X_{i,y};\boldsymbol{\theta}), y), \tag{1}$$

where l is the cross-entropy loss for classification problem. The transformations are defined as rotations by multiples of 90 degrees, i.e. $K = 4$, and $g(X;y)$ means rotating image X counterclockwise by $(y-1)\cdot 90$ degree.

The basic premise of RotNet is that rotating an image will change the orientation of objects in the image, which should be easily identified. To predict image rotation, the neural network has to recognize and localize salient object parts in the image. The well-trained neural network can therefore produce an accurate feature for salient object in the image and these features can be easily transferred to real-world tasks, such as detection and segmentation.

3.2. Noisy rotated images

The prerequisite introduced in the rotation prediction model could be satisfied for most natural images, which generally have objects in an up-front posture. This kind of images usually have a default orientation. Any rotations of the image will result in an unusual object orientation, which can be specified by human eyes without any doubt. Many instances in datasets like ImageNet have such observations, and are appropriate for the rotation prediction task.

Despite of its simplicity and effectiveness, this premise will fail for many objects in images that are orientation agnostic, for instance some objects viewed from top or in symmetrical shape (see Figure 1). Recognizing the exact rotation transformation for these images would be meaningless in practice, and applying ConvNets in any case without thinking will only introduce confounders to the model training. Moreover, features learned in RotNet are discriminative toward the rotation angle. They are not favored in a rotation agnostic image dataset like plankton [8] and ISBI 2012 electron microscopy segmentation challenge [3]. Here we first describe ways to reduce the influence of noisy rotation labels and introduce learning rotation unrelated features in the next subsection.

We regard original images in the dataset as being in default orientation and label them as positive examples. The unlabeled examples include all rotated copies, some of which are still in the default orientation after rotation. Therefore, the automatically assigned rotation labels of these images are noisy for RotNet. Predicting whether an input image is rotated is hence a binary classification problem if all unlabeled data are treated as negative examples [4]. In PU learning, it is shown that the estimated conditional probability is related to the noise rate and the confidence of an example being clean [40, 19]. We propose to weight each rotated image using estimated probability, and reduce the relative loss of rotation ambiguous images.

At first, a ConvNet model is trained to conduct binary classification. We denote by $\tilde{F}(X_{i,y})$ the probability of an image being positive estimated from this pre-trained model. We add a weight for each instance to the cross-entropy loss with tunable parameter γ, i.e.

$$w_{i,y} = \begin{cases} 1 & y = 1 \\ 1 - \tilde{F}(X_{i,y})^{\gamma} & \text{otherwise.} \end{cases} \tag{2}$$

The objective (1) can be reformulated using the calculated instance weights,

$$\min_{\boldsymbol{\theta}} \frac{1}{NK} \sum_{i=1}^{N} \sum_{y=1}^{K} w_{i,y} l(F(X_{i,y}; \boldsymbol{\theta}), y), \tag{3}$$

which predicts image rotations while mitigating the influence of noisy examples.

3.3. Feature decoupling

Image features that solely relate to image rotations are not practical for downstream tasks involving rotation agnostic images. An alternative solution is to complement rotation related feature with additional feature that is unrelated to image rotations. We achieve this goal by developing a feature decoupling algorithm, which learns a semantic feature that is partly discriminative with respect to image rotations and partly unrelated to it. The first part of the feature enjoys the benefits inherited from the task of estimating image rotations. Being unrelated to image rotations, the other part is suitable for some orientation agnostic tasks.

Rotation classification. We suppose that the high-level feature of an image X can be represented as $\mathbf{f} = \left[\mathbf{f}^{(1)\intercal}, \mathbf{f}^{(2)\intercal}\right]^{\intercal}$, where $\mathbf{f}^{(1)}$ is explicitly related to image rotation while $\mathbf{f}^{(2)}$ is responsible for information that are unrelated to rotation transformation. We denote by the ConvNet based feature extractor $F_f(\cdot; \boldsymbol{\theta}_f)$ with parameters $\boldsymbol{\theta}_f$, which maps an input rotated image $X_{i,y}$ into a fixed size vector $\mathbf{f}_{i,y} = F_f(X_{i,y}; \boldsymbol{\theta}_f)$. A classifier $F_c(\cdot; \boldsymbol{\theta}_c)$ takes feature $\mathbf{f}_{i,y}^{(1)}$ as the input to estimate the rotation type of the image. The rotation classication loss function can be expressed as

$$\mathcal{L}_c = \frac{1}{NK} \sum_{i=1}^{N} \sum_{y=1}^{K} w_{i,y} l(F_c(\mathbf{f}_{i,y}^{(1)}; \boldsymbol{\theta}_c), y), \tag{4}$$

which is different from Eq. (3) as only part of feature $\mathbf{f}$ are used here to recognize the rotation.

Rotation irrelevance. Toward the goal of rotation unrelated feature, we enforce similarity between features of the same image with different rotation angles. Formally, given rotated copies of an image: $\{X_y\}_{y=1}^{K}$, their features $\{\mathbf{f}_y^{(2)}\}_{y=1}^{K}$ are expected to be similar with each other as much as possible. We address this by minimizing the distance between each feature $\{\mathbf{f}_y^{(2)}\}_{y=1}^{K}$ and their mean feature vector $\bar{\mathbf{f}} = \frac{1}{K} \sum_{y=1}^{K} \mathbf{f}_y^{(2)}$, and write the objective as

$$\mathcal{L}_r = \frac{1}{NK} \sum_{i=1}^{N} \sum_{y=1}^{K} d(\mathbf{f}_{i,y}^{(2)}, \bar{\mathbf{f}}_i). \tag{5}$$

For calculation efficiency, we adopt Euclidean distance, i.e. $d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_2$.

However, this objective alone will lead to a trivial solution. Although features of the same image under different rotations are similar, the network could simply output the same vector (*e.g.* zero vector) regardless of the input image. Hence beyond Eq. (5), we expect rotation unrelated features to be discriminative w.r.t. image instances rather than rotation types. Non-parametric classification [48] can be applied to avoid such degenerate solution.

Image instance classification. The feature $\mathbf{f}^{(2)}$ is expected to be more similar with each other for the same image under different rotations than those for different images. Since features of rotated copies of an image have already been constrained to be close to their mean feature vector in Eq. (5), we proceed to distinguish and spread out these mean features.

In non-parametric classification, the probability of predicting image X as the i-th instance in the dataset is:

$$P(i \mid \hat{\mathbf{f}}) = \frac{\exp(\hat{\mathbf{f}}_i^{\intercal} \hat{\mathbf{f}} / \tau)}{\sum_{j=1}^{N} \exp(\hat{\mathbf{f}}_j^{\intercal} \hat{\mathbf{f}} / \tau)}, \tag{6}$$

where $\hat{\mathbf{f}}$ is the L2-normalized version of $\bar{\mathbf{f}}$ and τ is the temperature parameter. Given the training dataset S, we are interested in minimizing the negative log-likelihood:

$$\mathcal{L}_n = -\sum_{i=1}^{N} \log P(i \mid \hat{\mathbf{f}}_i). \tag{7}$$

To alleviate the time and space in demand to calculate Eq. (7) over large scale datasets, we linearly map the mean feature to a 128-dimentional vector before normalization as well as adopt noise constative estimation (NCE) and proximal regularization [48]. The objective is to minimize the following loss function:

$$\begin{aligned}\mathcal{L}_n = &-\mathbb{E}_{P_d}\left[\log h(i, \hat{\mathbf{f}}_i^{(t-1)}) - \lambda \left\|\hat{\mathbf{f}}_i^{(t)} - \hat{\mathbf{f}}_i^{(t-1)}\right\|_2^2\right] \\ &- m \cdot \mathbb{E}_{P_n}\left[\log(1 - h\big(i, \hat{\mathbf{f}}'^{(t-1)})\big)\right],\end{aligned} \tag{8}$$

where $h(i, \hat{\mathbf{f}}) := P(i \mid \hat{\mathbf{f}}) \big/ \big[P(i \mid \hat{\mathbf{f}}) + mP_n(i)\big]$. P_d denotes the actual data distribution and P_n denotes the uniform distribution for noise in NCE. $\hat{\mathbf{f}}'$ is the normalized feature from another image.

The resulting model comprises three core modules: rotation classification (Eq. (4)), rotation irrelevance (Eq. (5)) and image instance classification (Eq. (8)), and can be written as

$$\min_{\boldsymbol{\theta}_f, \boldsymbol{\theta}_c} \lambda_c \mathcal{L}_c + \lambda_r \mathcal{L}_r + \lambda_n \mathcal{L}_n. \tag{9}$$

We concatenate $\mathbf{f}^{(1)}$ and $\mathbf{f}^{(2)}$ to represent an input image. The image feature consists of rotation related and unrelated components, both of which contain rich high-level sematic image representation. $\mathbf{f}^{(1)}$ will contain necessary information, for example salient object location and its default orientation, to predict image rotations. On the other hand, $\mathbf{f}^{(2)}$ has no information related to rotation and focuses more on the differences of every single images.

4. Experiments

In this section, we conduct experiments to demonstrate the effectiveness of our approach. If the visual representations learned in unsupervised manner are effective and general-purpose, they will generalize well to various tasks. We first qualitatively analyze the network learned with the proposed algorithm. Then we report the results on several standard transfer learning benchmarks.

4.1. Implementation details

For comparison with previous works, we use a standard AlexNet architecture implemented by pytorch [24] with reduced number of channels as the feature extractor $F_f(\cdot; \boldsymbol{\theta}_f)$. It consists of five convolutional layers and two fully connected layers. We leave out the Local Response Normalization (LRN) layers and add Batch Normalization (BN) after each linear layers, which is a common procedure in recent self-supervised learning approaches [10, 52, 12, 53, 17, 48, 44, 6]. The decoupled features $\mathbf{f}^{(1)}$ and $\mathbf{f}^{(2)}$ are simply set to have same dimension, i.e. the representation $\mathbf{f}$ is split into two halves. We implement the rotation classifier $F_c(\cdot; \boldsymbol{\theta}_c)$ as a one-layer linear network. The value of hyperparameters γ, τ and m in our model are 2, 0.07 and 4096, respectively. We simply set the parameters λ_c, λ_r and λ_n for loss balance all as 1. We train the model for 200 epochs in total on ILSVRC 2012 training set. The learning rate is set to 0.01 initially and then decayed by a factor of 10 every 40 epochs after the first 90 epochs. The network is trained with momentum of 0.9, a batch size of 192 and an l_2 penalization of the weights $\boldsymbol{\theta}$ with $5 \cdot 10^{-4}$.

4.2. Qualitative analysis

Nearest-neighbor retrieval. Self-supervised training is expected to assign similar features to semantically similar images. We first perform nearest-neighbor retrieval on ImageNet ILSVRC 2012 validation set to test the ability of learned features in capturing semantic meanings. We compare to the RotNet baseline to see the effect of feature decoupling. For our model, we obtain features from the 4,096 dimensional vector outputted by the feature extractor network $F_f(\cdot; \boldsymbol{\theta}_f)$. Accordingly, for RotNet the features are extracted from the `fc7` layer. We use cosine-similarity to calculate the distance between features.

Retrievals of some examples are arranged from left to right in order of increasing distance in Figure 3. Both RotNet and the proposed model are able to capture semantics in images for some categories. The results of randomly selected images, which contain salient objects and are rotation unambiguous, are satisfactory for both RotNet and our model. Our model can sometimes capture more fine-grained similarity. For example, on second row, the RotNet retrieves some similar background plants rather than the foreground object bird. For bullet train, our model successfully find images in the same category rather than just general vehicles. Additionally, for some rotation agnostic image queries, the RotNet fail to extract latent information for objects in images. Many images retrieved by RotNet are totally unrelated to the query (marked with red border). This is likely because the RotNet focuses more on the shape of object and is less discriminative toward different instances. On the contrary, our model can return more semantically similar images for these queries, which confirms our model's discriminative ability on instance level.

Filter visualizations. To better understand the filers and features learned in our approach, we use different network visualization techniques. Figure 4 shows the filters from the first layer [25], synthetic images that maximize some activation [15, 50] and maximally activating images [51] for some channel of each convolutional layers. We find that deeper layers in our model seem to capture more complex and abstract textural structures.

4.3. Linear classification on activations

Following Zhang *et al.* [52], we train linear classifiers on top of the features extracted by different convolutional layers. This classification result represents the task spe-

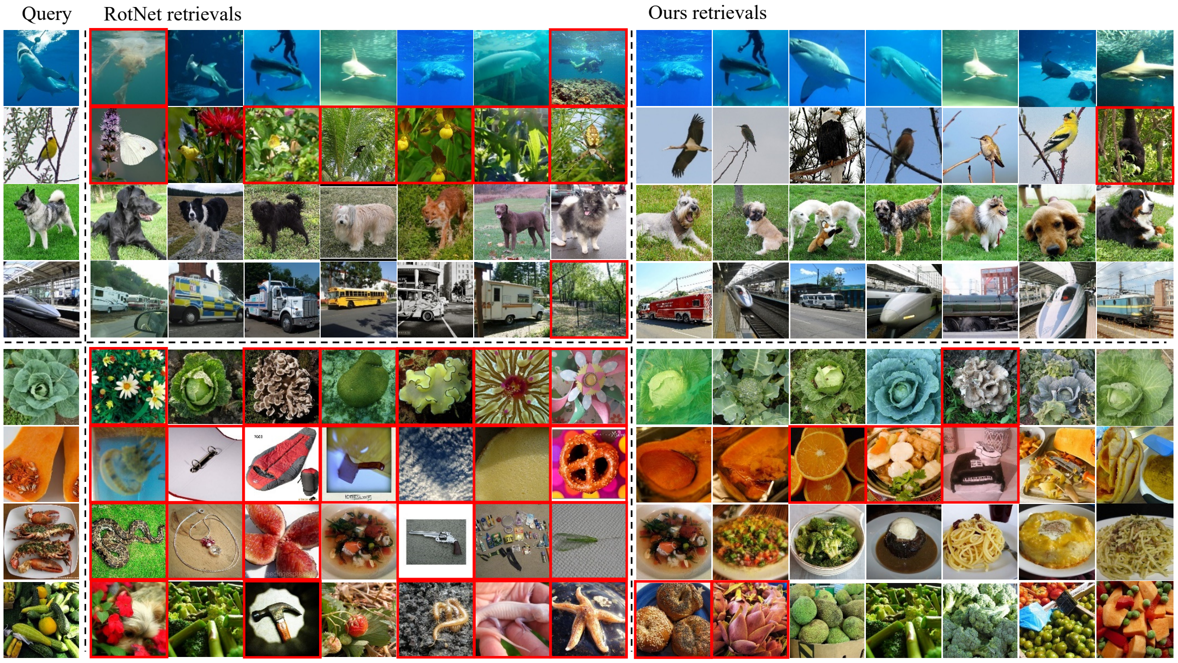
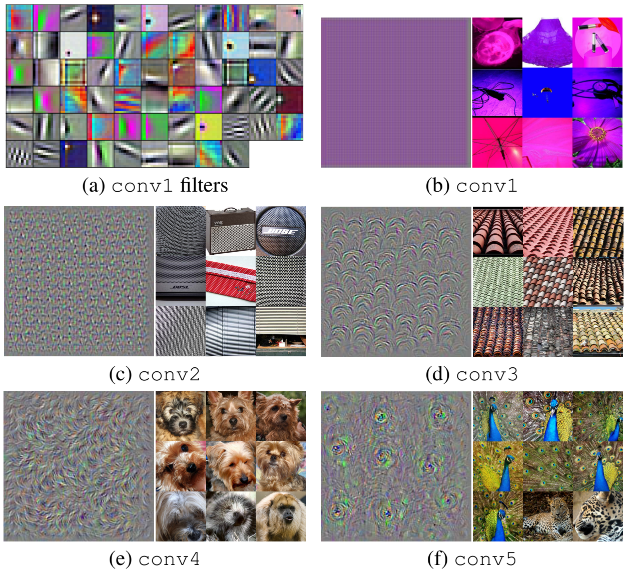


Figure 3: Nearest-neighbor retrieval results. We show the seven nearest neighbors of RotNet and our feature decoupling network on ImageNet validation set. Queries contain both randomly selected images (upper four rows) and rotation agnostic ones (lower four rows). Semantically unrelated retrievals are marked with red border.



(a) conv1 filters (b) conv1

(c) conv2 (d) conv3

(e) conv4 (f) conv5

Figure 4: Filter visualization. We plot the filters from conv1 layer and show synthetic images that maximally activate specific feature map of some channel in different convolutional layers. Corresponding top 9 activated images from ImageNet training set of that channel are on the right.

cific power of the learned representation, specifically the discriminative power over object class. As usual, we perform this study on both ILSVRC 2012 [45] and the Places dataset [54]. All weights of the feature extractor network are frozen and feature maps are spatially resized (with adaptive max pooling) so as to have around 9,000 elements. Results are reported in Tables 1 and 2, respectively. All approaches in table use AlexNet based network and were pretrained on ImageNet without labels except the ImageNet-labels, Places-labels, and Random entries.[1] We report the best numbers for each method reported in [36, 39]. We also provide results of non-linear classification on ImageNet in the supplementary material.

On ImageNet, our approach outperforms the state-of-the-art from conv3 to conv5. Our results on conv1 and conv2 are comparable to previous results and the ImageNet-label entry. Note that the lower layers of the network usually capture low-level information like edges or contours in images, and with relatively low transfer accuracies, these features are generally less often used directly. It is important to note that the performance of most previous works degrades along the neural network depth. In stark contrast, we successfully diminish the gap with ImageNet-label on higher layers. The largest improvement (7.8%) is achieved on conv5 layer, which usually extracts abstract semantic information. This suggests that high-level features extracted by our method are more promising for practical use.

On Places dataset, the results of our approach exhibit similar trends with that on ImageNet. We achieve best accu-

[1] Methods marked with * use a *bigger* version of AlexNet that do not have *group* or reduced number of channels, which will have 50% more parameters in convolutional layers and typically cause a performance boost. In this paper we also report results on this network.

Method\Layer	conv1	conv2	conv3	conv4	conv5
ImageNet-labels [25, 52]	19.3	36.3	44.2	48.3	50.5
Random [53]	11.6	17.1	16.9	16.3	14.1
Krähenbühl *et al*. [22]	17.5	23.0	24.5	23.2	20.6
Pathak *et al*. (Inpainting) [43]	14.1	20.7	21.0	19.8	15.5
Noroozi & Favaro (Jigsaw) [37]	18.2	28.8	34.0	33.9	27.1
Zhang *et al*. (Colorization) [52]	13.1	24.8	31.0	32.6	31.8
Donahue *et al*. (BiGANs) [12]	17.7	24.5	31.0	29.9	28.0
Zhang *et al*. (Split-Brain) [53]	17.7	29.3	35.4	35.2	32.8
Noroozi *et al*. (Counting) [38]	18.0	30.6	34.3	32.5	25.7
Gidaris *et al*. (RotNet) [17]	18.8	31.7	38.7	38.2	36.5
Jenni & Favaro [21]	19.5	**33.3**	37.9	38.9	34.9
Mundhenk *et al*. [36]	**19.6**	31.8	37.6	37.8	33.7
Noroozi *et al*. (CC+) [39]	18.9	30.5	35.7	35.4	32.2
Noroozi *et al*. (CC+vgg-) [39]	19.2	32.0	37.3	37.1	34.6
Wu *et al*. [48]	16.8	26.5	31.8	34.1	35.6
Doersch *et al*. (Context) [10]*	16.2	23.3	30.2	31.7	29.6
Ren & Lee [44]*	16.5	27.0	30.5	30.1	26.5
Caron *et al*. (DeepCluster) [6]*†	13.4	32.3	41.0	39.6	38.2
Ours	19.3	**33.3**	**40.8**	**41.8**	**44.3**
Ours (*bigger* AlexNet)*	20.8	35.2	41.8	44.3	44.4
Ours (*bigger* AlexNet)*†	22.2	38.2	45.7	48.7	48.3

Table 1: Top-1 linear classification accuracies on ImageNet validation set using activations from different convolutional layers as features. * indicates the use of a *bigger* AlexNet. † indicates reporting accuracies averaged over 10 crops.

Method\Layer	conv1	conv2	conv3	conv4	conv5
Places-labels [54, 53]	22.1	35.1	40.2	43.3	44.6
ImageNet-labels [25, 52]	22.7	34.8	38.4	39.4	38.7
Random [53]	15.7	20.3	19.8	19.1	17.5
Krähenbühl *et al*. [22]	21.4	26.2	27.1	26.1	24.0
Pathak *et al*. (Inpainting) [43]	18.2	23.2	23.4	21.9	18.4
Noroozi & Favaro (Jigsaw) [37]	23.0	31.9	35.0	34.2	29.3
Zhang *et al*. (Colorization) [52]	16.0	25.7	29.6	30.3	29.7
Donahue *et al*. (BiGANs) [12]	22.0	28.7	31.8	31.3	29.7
Zhang *et al*. (Split-Brain) [53]	21.3	30.7	34.0	34.1	32.5
Noroozi *et al*. (Counting) [38]	23.3	33.9	36.3	34.7	29.6
Gidaris *et al*. (RotNet) [17]	21.5	31.0	35.1	34.6	33.7
Jenni & Favaro [21]	23.3	**34.3**	36.9	**37.3**	34.4
Mundhenk *et al*. [36]	**23.7**	34.2	37.2	37.2	34.9
Noroozi *et al*. (CC+) [39]	22.5	33.0	36.2	36.1	34.0
Noroozi *et al*. (CC+vgg-) [39]	22.9	34.2	**37.5**	37.1	34.4
Wu *et al*. [48]	18.8	24.3	31.9	34.5	33.6
Doersch *et al*. (Context) [10]*	19.7	26.7	31.9	32.7	30.9
Caron *et al*. (DeepCluster) [6]*†	19.6	33.2	39.2	39.8	34.7
Ours	22.9	32.4	36.6	**37.3**	**38.6**
Ours (*bigger* AlexNet)*	24.0	33.8	37.5	39.3	38.9
Ours (*bigger* AlexNet)*†	25.5	36.0	40.1	42.2	41.3

Table 2: Top-1 linear classification accuracies on Places validation set using activations from different convolutional layers as features. * indicates the use of a *bigger* AlexNet. † indicates reporting accuracies averaged over 10 crops.

racies on `conv4` and `conv5` layer, as well as comparable accuracies from `conv1` to `conv3`. On `conv5` layer we outperform the state-of-the-art by 3.7%.

4.4. Multi-label classification, object detection and semantic segmentation on PASCAL VOC

We test the transferability of the learned feature on PASCAL VOC dataset [16]. We use our unsupervised trained network $F_f(\cdot;\boldsymbol{\theta}_f)$ as the initialization model for tasks on PASCAL. Performance is measured by mean average precision (mAP) for classification and detection, and by mean intersection over union (mIU) for segmentation. During transfer, we absorb the batch normalization parameters into their preceding linear layers and do not use BN layers during fine-tuning. The data-dependent rescaling method proposed by Krähenbühl *et al.* [22] is used to rescale the weights in all experiments as is standard practice. Table 3 summarizes the comparison of our approach with other methods. We outperform previous methods on all these three tasks.

Classification on PASCAL VOC 2007. We use the open source protocol provided by Krähenbühl [2] to perform multi-label classification. We fine-tune either the whole network or only `fc6-8` layers on *trainval* set and evaluate on *test* set. Our approach can improve upon RotNet, the current best method on classification. It can be observed that the *bigger* AlexNet model will lead to a performance improvement.

Detection on PASCAL VOC 2007. For object detection our self-supervised trained network is used as the initialization of Fast-RCNN [18]. We use the publicly available testing framework provided by Girshick [18] and use multi-scale training and single-scale testing. The weights of the first layer are fixed during fine-tuning as it is the default setting in Fast-RCNN. With a mAP of 57.5% we achieve the best result. Per class detection performance of our method is also provided in the supplementary material.

Segmentation on PASCAL VOC 2012. We fine-tune our model using FCN [33] on PASCAL VOC 2012 *train* set and evaluate on *val* set. Our approach outperforms state-of-the-art by 2.7%.

4.5. Discussion

Ablation studies. To see the influence of each component in our model, we conduct ablation studies on ImageNet linear classification with fixed features. We compare the individual performance of the rotation prediction task (Rotation), rotation unrelated instance classification (Instance), the combination of these two tasks (Rotation + Instance), and the full model taking into consideration noisy labels in the unlabeled set (PURotation + Instance). The middle four rows in Table 4 shows the resulting performance of different components. The model performs best when rotation

[2] https://github.com/philkr/voc-classification

Method\Task	Class.		Det.	Seg.
	fc6-8	all	all	all
ImageNet-labels [25, 52, 43]	78.9	79.9	59.1 [39]	48.0
Random [43]	–	53.3	43.4	–
Autoencoder [12]	–	53.8	41.9	–
Krähenbühl *et al.* [22]	39.2	56.6	45.6	32.6
Pathak *et al.* (Inpainting) [43]	34.6	56.5	44.5	29.7
Noroozi & Favaro (Jigsaw) [37]	–	67.6	53.2	37.6
Zhang *et al.* (Colorization) [52]	61.5	65.6	46.9	35.6
Donahue *et al.* (BiGANs) [12]	52.3	60.1	46.9	35.2
Larsson *et al.* (Colorization) [28]	–	65.9	–	38.4
Zhang *et al.* (Split-Brain) [53]	63.0	67.1	46.7	36.0
Noroozi *et al.* (Counting) [38]	–	67.7	51.4	36.6
Gidaris *et al.* (RotNet) [17]	70.9	73.0	54.4	39.1
Jenni & Favaro [21]	–	69.8	52.5	38.1
Mundhenk *et al.* [36]	–	69.6	55.8	41.4
Noroozi *et al.* (CC+) [39]	–	69.9	55.0	40.0
Noroozi *et al.* (CC+vgg-) [39]	–	72.5	56.5	42.6
Wu *et al.* [48]	–	–	48.1	–
Doersch *et al.* (Context) [10]*	55.1	65.3	51.1	–
Ren & Lee [44]*	–	68.0	52.6	–
Caron *et al.* (DeepCluster) [6]*	72.0	73.7	55.4	45.1
Ours	**72.3**	**74.3**	**57.5**	**45.3**
Ours (*bigger* AlexNet)*	72.5	74.7	58.0	45.9

Table 3: Transfer learning results for classification, detection and segmentation on PASCAL compared to state-of-the-art feature learning methods. We report the best numbers for each method reported in [36, 39]. * indicates the use of a *bigger* AlexNet.

discrimination, noisy labels and instance discrimination are all considered.

Different configurations. We evaluate the effect of various design choices by linear classification on ImageNet. We compare different structures of the feature extractor network $F_f(\cdot;\boldsymbol{\theta}_f)$: the convolutional layers of AlexNet (conv5), conv5 with one fully connected layer (fc6), and conv5 with two fully connected layers (fc7). Results are summarized at the lower three rows in Table 4. Higher layers learn better feature when feature decouples at higher layers. It is interesting to notice that performance of lower layers tend to decrease. This might because effective gradient information help less on lower layers when the loss function is applied on higher layers.

Rotation feature evaluation. We finally demonstrate that the decoupled feature is better suitable when the images in downstream tasks exhibit rotational symmetry. To do this, we rotate the images in PASCAL VOC 2007 by multiples of 90 degrees (specifically 90, 180 and 270) and evaluate on classification task. The rotated dataset has 20,044 images for training and 19,808 for test (4 times as many as the original dataset). Each instance with different rotation angles shares the same class label. We train a linear classifier directly on top of the first half (rotation related) features $\mathbf{f}^{(1)}$, second half (rotation unrelated) features $\mathbf{f}^{(2)}$ and the compound decoupled features. Results for ImageNet-labels and baseline RotNet are produced by us and are reported for reference. We also consider fine-tuning fc6-8 for three different methods. As shown in Table 5, the learned feature of RotNet and our rotation related half performs poorly. The reason is that they are discriminative w.r.t. image rotations and do not have a good generalization ability in a rotated dataset. This result reveals that it is beneficial to consider both rotation related and unrelated features. Our method is more suitable in vision tasks that are in favor of rotation invariance.

Method	Decouple	conv1	conv2	conv3	conv4	conv5
ImageNet-labels	–	19.3	36.3	44.2	48.3	50.5
Rotation	–	18.8	31.7	38.7	38.2	36.5
Instance	–	18.3	28.6	33.0	32.7	32.9
Rotation + Instance	fc7	19.3	33.0	40.7	41.6	44.0
PURotation + Instance (Full model)	fc7	19.3	33.3	40.8	41.8	44.3
Full model	conv5	19.6	33.4	40.2	40.4	41.0
Full model	fc6	19.4	33.5	40.8	41.5	42.6
Full model	fc7	19.3	33.3	40.8	41.8	44.3

Table 4: Comparison of different components and design choices in our model on ImageNet linear classification task.

Method\Task	Class. (fc8)	Class. (fc6-8)
ImageNet-labels	66.5	71.6
RotNet	42.2	66.3
Ours (rotation related half $\mathbf{f}^{(1)}$)	38.6	–
Ours (rotation unrelated half $\mathbf{f}^{(2)}$)	57.7	–
Ours (decoupled feature $\mathbf{f}$)	59.2	68.0

Table 5: Rotation feature evaluation results on Rotated PASCAL classification.

5. Conclusion

In this paper, we have presented an unsupervised representation learning method that learns semantically meaningful features containing rotation related and unrelated parts. Our approach decouples predicting image rotations from discriminating individual instances. The transfer of features achieves improved performance over state-of-the-art methods on standard self-supervised learning benchmarks. The advantages of decoupled feature are further demonstrated in rotation agnostic tasks. We believe that incorporating more well-analyzed properties of representation for self-supervised learning is beneficial to generalization and is a promising future direction.

Acknowledgement

This work was supported in part by the Australian Research Council under Project FL-170100117, Project DP-180103424, Project IH180100002, and Project DE-180101438.

References

[1] Pulkit Agrawal, Joao Carreira, and Jitendra Malik. Learning to see by moving. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[3] Ignacio Arganda-Carreras, Srinivas C. Turaga, Daniel R. Berger, Dan Cireşan, Alessandro Giusti, Luca M. Gambardella, Jürgen Schmidhuber, Dmitry Laptev, Sarvesh Dwivedi, Joachim M. Buhmann, Ting Liu, Mojtaba Seyedhosseini, Tolga Tasdizen, Lee Kamentsky, Radim Burget, Vaclav Uher, Xiao Tan, Changming Sun, Tuan D. Pham, Erhan Bas, Mustafa G. Uzunbas, Albert Cardona, Johannes Schindelin, and H. Sebastian Seung. Crowdsourcing the creation of image segmentation algorithms for connectomics. *Frontiers in Neuroanatomy*, 9:142, 2015.

[4] Jessa Bekker and Jesse Davis. Learning from positive and unlabeled data: A survey. *arXiv:1811.04820*, 2018.

[5] Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, Aug 2013.

[6] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision – ECCV 2018*, pages 139–156, Cham, 2018. Springer International Publishing.

[7] Taco Cohen and Max Welling. Group equivariant convolutional networks. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 2990–2999, New York, New York, USA, 20–22 Jun 2016. PMLR.

[8] Robert K Cowen, S Sponaugle, K Robinson, and J Luo. Planktonset 1.0: Plankton imagery data collected from fg walton smith in straits of florida from 2014–06-03 to 2014–06-06 and used in the 2015 national data science bowl (ncei accession 0127422). *NOAA National Centers for Environmental Information*, 2015.

[9] Sander Dieleman, Jeffrey De Fauw, and Koray Kavukcuoglu. Exploiting cyclic symmetry in convolutional neural networks. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1889–1898, New York, New York, USA, 20–22 Jun 2016. PMLR.

[10] Carl Doersch, Abhinav Gupta, and Alexei A. Efros. Unsupervised visual representation learning by context prediction. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[11] Carl Doersch and Andrew Zisserman. Multi-task self-supervised visual learning. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[12] Jeff Donahue, Philipp Krähenbühl, and Trevor Darrell. Adversarial feature learning. In *International Conference on Learning Representations*, 2017.

[13] Alexey Dosovitskiy, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. Discriminative unsupervised feature learning with convolutional neural networks. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 766–774. Curran Associates, Inc., 2014.

[14] Charles Elkan and Keith Noto. Learning classifiers from only positive and unlabeled data. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, pages 213–220, New York, NY, USA, 2008. ACM.

[15] Dumitru Erhan, Yoshua Bengio, Aaron Courville, and Pascal Vincent. Visualizing higher-layer features of a deep network. Technical Report 1341, University of Montreal, June 2009. Also presented at the ICML 2009 Workshop on Learning Feature Hierarchies, Montréal, Canada.

[16] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111(1):98–136, Jan. 2015.

[17] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *International Conference on Learning Representations*, 2018.

[18] Ross Girshick. Fast r-cnn. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[19] Fengxiang He, Tongliang Liu, Geoffrey I Webb, and Dacheng Tao. Instance-dependent PU learning by bayesian optimal relabeling. *arXiv:1808.02180*, 2018.

[20] João F. Henriques and Andrea Vedaldi. Warped convolutions: Efficient invariance to spatial transformations. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1461–1469, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR.

[21] Simon Jenni and Paolo Favaro. Self-supervised feature learning by learning to spot artifacts. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[22] Philipp Krähenbühl, Carl Doersch, Jeff Donahue, and Trevor Darrell. Data-dependent initializations of convolutional neural networks. In *International Conference on Learning Representations*, 2016.

[23] Jan Kremer, Fei Sha, and Christian Igel. Robust active label correction. In Amos Storkey and Fernando Perez-Cruz, editors, *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 308–316, Playa Blanca, Lanzarote, Canary Islands, 09–11 Apr 2018. PMLR.

[24] Alex Krizhevsky. One weird trick for parallelizing convolutional neural networks. *CoRR*, abs/1404.5997, 2014.

[25] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

[26] Dmitry Laptev, Nikolay Savinov, Joachim M. Buhmann, and Marc Pollefeys. Ti-pooling: Transformation-invariant pooling for feature learning in convolutional neural networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[27] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Learning representations for automatic colorization. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 577–593, Cham, 2016. Springer International Publishing.

[28] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Colorization as a proxy task for visual understanding. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[29] S. Lazebnik, C. Schmid, and Jean Ponce. Semi-local affine parts for object recognition. In *Proceedings of the British Machine Vision Conference*, pages 98.1–98.10. BMVA Press, 2004. doi:10.5244/C.18.98.

[30] Hsin-Ying Lee, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Unsupervised representation learning by sorting sequences. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[31] 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 David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 740–755, Cham, 2014. Springer International Publishing.

[32] T. Liu and D. Tao. Classification with noisy labels by importance reweighting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(3):447–461, March 2016.

[33] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[34] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, Nov 2004.

[35] Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. Learning with noisy labels. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 1196–1204. Curran Associates, Inc., 2013.

[36] T. Nathan Mundhenk, Daniel Ho, and Barry Y. Chen. Improvements to context based self-supervised learning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[37] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 69–84, Cham, 2016. Springer International Publishing.

[38] Mehdi Noroozi, Hamed Pirsiavash, and Paolo Favaro. Representation learning by learning to count. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[39] Mehdi Noroozi, Ananth Vinjimoor, Paolo Favaro, and Hamed Pirsiavash. Boosting self-supervised learning via knowledge transfer. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[40] Curtis G. Northcutt, Tailin Wu, and Isaac L. Chuang. Learning with confident examples: Rank pruning for robust classification with noisy labels. In *Proceedings of the Thirty-Third Conference on Uncertainty in Artificial Intelligence*, UAI'17. AUAI Press, 2017.

[41] Andrew Owens, Jiajun Wu, Josh H. McDermott, William T. Freeman, and Antonio Torralba. Ambient sound provides supervision for visual learning. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 801–816, Cham, 2016. Springer International Publishing.

[42] Deepak Pathak, Ross Girshick, Piotr Dollar, Trevor Darrell, and Bharath Hariharan. Learning features by watching objects move. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[43] Deepak Pathak, Philipp Krähenbühl, Jeff Donahue, Trevor Darrell, and Alexei A. Efros. Context encoders: Feature learning by inpainting. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[44] Zhongzheng Ren and Yong Jae Lee. Cross-domain self-supervised multi-task feature learning using synthetic imagery. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[45] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, Dec 2015.

[46] Clayton Scott. A Rate of Convergence for Mixture Proportion Estimation, with Application to Learning from Noisy Labels. In Guy Lebanon and S. V. N. Vishwanathan, editors, *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, volume 38 of *Proceedings of Machine Learning Research*, pages 838–846, San Diego, California, USA, 09–12 May 2015. PMLR.

[47] Xiaolong Wang and Abhinav Gupta. Unsupervised learning of visual representations using videos. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[48] Zhirong Wu, Yuanjun Xiong, Stella X. Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[49] Pengyi Yang, Wei Liu, and Jean Yang. Positive unlabeled learning via wrapper-based adaptive sampling. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 3273–3279, 2017.

[50] Jason Yosinski, Jeff Clune, Thomas Fuchs, and Hod Lipson. Understanding neural networks through deep visualization. In *ICML Workshop on Deep Learning*, 2015.

[51] Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 818–833, Cham, 2014. Springer International Publishing.

[52] Richard Zhang, Phillip Isola, and Alexei A. Efros. Colorful image colorization. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 649–666, Cham, 2016. Springer International Publishing.

[53] Richard Zhang, Phillip Isola, and Alexei A. Efros. Split-brain autoencoders: Unsupervised learning by cross-channel prediction. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[54] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using places database. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 487–495. Curran Associates, Inc., 2014.