NOVELTY DETECTION VIA ROTATED CONTRASTIVE PREDICTIVE CODING

Anonymous authors
Paper under double-blind review

ABSTRACT

The current dominant paradigm for novelty detection relies on a learned model's capability to *recover* the regularities. To this end, reconstruction-based learning is often used in which the normality of an observation is expressed in how well it can be reconstructed. However, this can be limiting as anomalous data can be reconstructed well if enough common features are shared between normal and anomalous data. In this paper, we pursue an alternative approach wherein the normality is measured by a contrastive learning objective. Specifically, we propose Rotated Contrastive Predictive Coding (Rotated CPC) where the model operates on rotated images and simultaneously learns to predict the future in latent space. Normality score is thus measured as how predictive the representations are and the score's robustness is further improved by ensembling predictions on multiple rotations of the input signal. We demonstrate the efficacy of this formulation across a variety of benchmark datasets where our method outperforms state-of-the-art methods.

1 INTRODUCTION

Novelty detection tackles the problem of identifying anomalous samples that deviate from the regularity defined by a model, which is learned from a collection of normal data. It is a problem that is useful in many real-world applications that require detecting unseen or surprising events, or knowing in advance when a system will malfunction; for instance, it has received significant attention in application areas such as user/intruder authentication (Garcia-Teodoro et al., 2009), medical imaging (Schlegl et al., 2017), video surveillance (Abati et al., 2019), and defect detection (Clifton et al., 2007).

There are a few possible variants for the novelty detection problem. There is a supervised scenario where both the normal and abnormal samples are provided during training. Either being fully supervised where labeled anomalous samples exist (Hendrycks et al., 2018) or being semi-supervised where a normal dataset and an unlabeled set contaminated with anomalous samples are provided (Blanchard et al., 2010), these variants assume some type of access to anomalies which may not always be possible. In this paper, we focus on the unsupervised variant of the problem where only inliers are available for training. Moreover, we address novelty detection problem for images where the high-dimensional nature of the data makes it more challenging.

A large body of contemporary work in novelty detection for vision primarily leverage, implicitly or explicitly, some form of reconstruction-based learning. Reconstruction-based methods can be categorized into mainly two strategies: the first strategy consists of approaches that formulate a novelty score by analyzing the deficiency in the reconstructed data point, mainly through deep autoencoders (AEs) or generative adversarial networks (GANs) (Xia et al., 2015; Schlegl et al., 2017). Various distance measures such as mean squared error between the query image and its reconstruction, or the discriminator output have been extensively explored for this purpose. The second strategy consists of methods that use reconstruction as a task to learn a representative latent space for normality (Pidhorskyi et al., 2018; Zong et al., 2018). The learned latent embeddings are then used to identify outliers by fitting a density model or learning a one-class classifier over these embeddings. The general assumption in these reconstruction-based strategies is that the anomaly incurs higher reconstruction error and thus how well an input can be remembered or recovered serves as a good indicator or learning signal for novelty detection.

However, this assumption does not hold at all times. Observations have been made in existing literature (Gong et al., 2019; Salehi et al., 2020; Perera et al., 2019; Zong et al., 2018) that methods can "generalize" too well when the model complexity is high or when inliers and outliers share common features or patterns. For instance, Perera et al. (2019) and Salehi et al. (2020) demonstrate that an autoencoder trained only on digits of 8 can provide good reconstruction for digits 1,5,6 and 9. Such generalization property is considered as a drawback for novelty detection systems that rely on reconstruction-based learning.

In this work, we take an alternative approach based on contrastive learning that does not require any reconstruction, but instead relies on solving a semantic pretext task to learn representations useful for modeling normality. Concretely, we train a powerful autoregressive model under the Contrastive Predictive Coding (CPC) framework (Oord et al., 2018b) and use Noise Contrastive Estimation objective (Gutmann & Hyvärinen, 2010; Mnih & Teh, 2012; Jozefowicz et al., 2016) as the novelty score. The model is trained by predicting the future given context in the latent space—a task that enforces the model to learn representations encoding the underlying shared information between patches of images while ignoring background cues and local information such as low-level noise. We postulate that this semantic task is not only beneficial for learning good representations, but also serves as a good mechanism to detect anomalous samples. The intuition is that the representations learned with normal data should be less predicative of the future for anomalous data. We further reinforce this idea by operating the CPC model on rotated images so that the model learns separate subspaces in which it optimizes for the contrastive objective.

We present extensive experiments of the proposed method on multiple benchmark datasets and demonstrate that our method outperforms reconstruction-based novelty detection systems and also achieves competitive or better performance compared to non reconstruction-based methods.

2 Related Work

The literature in novelty detection is quite massive, so we refer to survey papers (Chalapathy & Chawla, 2019; Zimek et al., 2012; Chandola et al., 2009) for a more comprehensive view. In this work, we mainly focus on the more recent approaches that leverage deep models to tackle novelty detection problems on high-dimensional data such as images.

Reconstruction-based Methods. Reconstruction-based methods are some of the most common approaches in novelty detection. The main idea that motivates these methods is that anomalous samples should incur high reconstruction error when using a model that has only seen normal samples. Many methods rely on deep autoencoders in which the reconstruction error is used to derive a novelty score (Xia et al., 2015; Abati et al., 2019; Gong et al., 2019; Salehi et al., 2020). Another set of recent methods use GANs to reconstruct normal samples (Schlegl et al., 2017; Perera et al., 2019) where either the mean squared error or the discriminator output is used as a measure of novelty. Other than computing the novelty score, reconstruction task can also be used as a way to learn latent representations on which a probabilistic model can be fitted. For instance, Pidhorskyi et al. (2018) and Zong et al. (2018) use Gaussian or Gaussian mixture models to estimate the density of the latents learned from an autoencoder. By jointly optimizing the autoencoder and the probabilistic model, both methods are able to regularize the representations and encourage better modeling of inlier distribution.

Non-Reconstruction-based Methods. There are also novelty detection approaches that do not rely on any form of reconstruction. Ruff et al. (2018) proposes a deep one-class SVM model that maps normal data into a hypersphere of minimum volume. Another popular paradigm for novelty detection recently has been based on self-supervised learning. For example, Golan & El-Yaniv (2018) and Bergman & Hoshen (2020) train a deep model with a pretext task of predicting what type of geometric transformation (i.e. horizontal flipping, translations, rotations, etc.) was applied to the image of interest. The generalization error on this pretext task is then used to detect novelty. Our proposed method falls in this group of methodologies, but differs in that it relies on a contrastive learning framework augmented with rotated images.

Contrastive Learning. Discriminative approaches based on contrastive learning in the latent space have recently shown great promise, achieving state-of-the-art results. (Oord et al., 2018a; He et al.,

2020; Chen et al., 2020; Grill et al., 2020; Bachman et al., 2019). By contrasting positive pairs and negative pairs, these methods try to learn useful visual representations that are transferable to other downstream tasks. In this work, we investigate how contrastive learning framework can be used in novelty detection setting. In particular, we take CPC as the basis of our method. It has been shown that CPC can be applied to various domains such as speech, image, text, and reinforcement learning; however, our work is the first to use CPC in the novelty detection setting.

3 Novelty Detection via Rotated Contrastive Predictive Coding

We start this section by defining the preliminaries of the problem (section 3.1), and reviewing the CPC architecture and its learning objective (section 3.2). Next, we introduce our method which we call Rotated Contrastive Predictive Coding (Rotated CPC) (section 3.3). After that we explain the novelty score function which is based on Noise-Contrastive Estimation and ensembling (section 3.4).

3.1 Problem Statement

Given a dataset $\boldsymbol{X} = \{x_1, x_2, ...x_n\}_{n=1}^{N}$ of N normal samples (i.e. $\boldsymbol{X} \sim p_{\text{normal}}(x)$), the goal of novelty detection is to learn a model only from $\boldsymbol{X}$ that detects whether a given sample belongs to the normal data distribution $p_{\text{normal}}(x)$.

It is a common practice to learn the scoring function $ns(x)$ as a proxy for modeling $p_{\text{normal}}(x)$ where higher scores indicate that x is more likely to be an inlier. Although the scoring function can be used to construct a binary decision boundary by thresholding the values, many existing works only focus on studying the trade-off between true positive rate and false positive rate–often measured by the area under the Receiver Operating Characteristic curve (AUROC). In this work, we also only focus on learning the scoring function and evaluate the proposed method using the AUROC metric.

3.2 Contrastive Predictive Coding

Contrastive Predictive Coding (Oord et al., 2018b) learns representations by training a powerful deep autoregressive model that predicts future observations given the past context in the latent space. The intuition behind this approach is to learn the underlying shared information between different parts of the data while ignoring low-level information that is rather local and noisy. By predicting far in the future instead in the vicinity, the encoded representations are forced to capture the shared global structure of the data.

In the case of visual representation learning, the input image $x \in \boldsymbol{X}$ is first divided into a grid of overlapping patches $x_{i,j}$ where i, j denote the location of the patch. A non-linear encoder g_{enc} maps each image patch to a latent representation $\boldsymbol{z}_{i,j} = g_{\text{enc}}(x_{i,j})$. The context feature $\boldsymbol{c}_{i,j}$ is then extracted through an autoregressive model g_{ar} which is a masked convolutional network (Oord et al., 2016) whose receptive field only covers features above the location i, j (i.e. $\boldsymbol{c}_{i,j} = g_{\text{ar}}(\boldsymbol{z}_{u,v})$ where $u \leq i$ and $v \leq j$ if $u = i$). In order to make predictions of future latent observations $\boldsymbol{z}_{i+k,j}$ where $k > 0$, a linear model is used: $\hat{\boldsymbol{z}}_{i+k,j} = \boldsymbol{W}_k \boldsymbol{c}_{i,j}$. The predicted feature vector $\hat{\boldsymbol{z}}_{i+k,j}$ is compared with the target feature vector $\boldsymbol{z}_{i+k,j}$ and other randomly sampled negative feature vectors $\{\boldsymbol{z}_l\}$. Specifically, the softmax operator is used to compute the probability of classifying the positive sample correctly:

$$p(\boldsymbol{z}_{i+k,j} | \hat{\boldsymbol{z}}_{i+k,j}, \{\boldsymbol{z}_l\}) = \frac{\exp\left(\langle \hat{\boldsymbol{z}}_{i+k,j}, \boldsymbol{z}_{i+k,j} \rangle\right)}{\exp\left(\langle \hat{\boldsymbol{z}}_{i+k,j}, \boldsymbol{z}_{i+k,j} \rangle\right) + \sum_l \exp\left(\langle \hat{\boldsymbol{z}}_{i+k,j}, \boldsymbol{z}_l \rangle\right)} \tag{1}$$

where $\langle , \rangle$ denotes the inner product.

The computed probability is then evaluated by the categorical cross-entropy loss:

$$\mathcal{L}_{CPC} = -\frac{1}{N} \sum_{x \in \boldsymbol{X}} \sum_{i,j,k} \log p(\boldsymbol{z}_{i+k,j} | \hat{\boldsymbol{z}}_{i+k,j}, \{\boldsymbol{z}_l\}) \tag{2}$$

The above objective is called InfoNCE (Oord et al., 2018b), which is a loss derived from Noise-Contrastive Estimation (Gutmann & Hyvärinen, 2010; Mnih & Teh, 2012; Jozefowicz et al., 2016),

and it has been shown that minimizing the InfoNCE loss $\mathcal{L}_{CPC}$ maximizes a lower bound on mutual information between $\boldsymbol{z}_n^{i+k,j}$ and $\boldsymbol{c}_n^{i,j}$. Both the encoder g_{enc} and the autoregressive model g_{ar} are jointly optimized under this objective.

3.3 ROTATED CPC

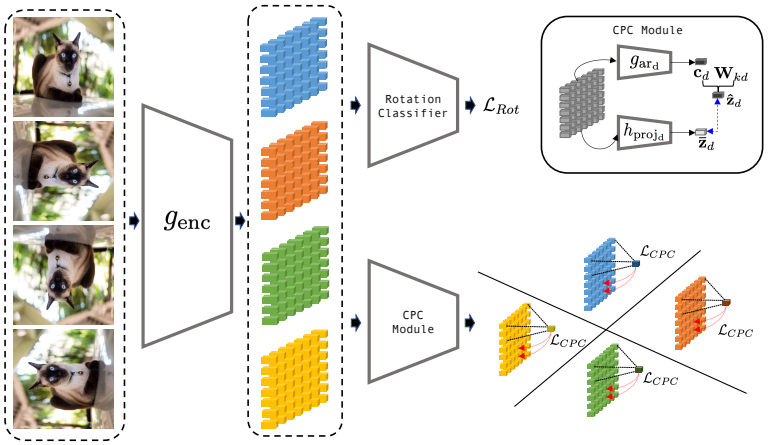

Figure 1: Rotated CPC consists of a patch-level encoder g_{enc}, a CPC module, and a rotation classifier. Within the CPC module, there are 4 autoregressive model and projection layer pairs for each prediction direction (i.e. $d \in \{\text{top-down, bottom-up, right-left, left-right}\}$). To optimize its objective on rotated images, the CPC module is forced to learn separate latent subspaces.

Hénaff et al. (2019) proposes an improved version (CPCv2) of the original CPC by making modifications in the encoder architecture, maximizing the supervisory signal obtained from each image, and introducing stochastic data-preprocessing techniques for individual patches. Concretely, CPCv2 achieves improved performance on downstream tasks by replacing Batch Normalization (Ioffe & Szegedy, 2015) with Layer Normalization (Ba et al., 2016) in the encoder, using additional non-linear projection layer h_{proj} on latents $\boldsymbol{z}_{i,j}$ (i.e. $\bar{\boldsymbol{z}}_{i,j} = h_{\text{proj}}(\boldsymbol{z}_{i,j})$). Thus, Equation 2 is modified by substituting all $\boldsymbol{z}$ with $\bar{\boldsymbol{z}}$. Additionally, CPCv2 predicts the future latent representations not only from top to bottom of the image, but in all four directions (top-down, bottom-up, left-right, right-left), and processes each patch using data augmentation schemes such as spatial/color jittering and color dropping (Doersch et al., 2015) in order to make the network avoid using low-level cues for solving the CPC objective. Our proposed Rotated CPC extends CPCv2 and repurposes it for novelty detection task.

Rotated CPC augments CPCv2 by creating additional subspaces in which the CPC objective is solved. We achieve this by having the model operate on rotated images.

Let $\mathcal{R}_\theta$ be a set of rotation operations where $\theta \in \{0°, 90°, 180°, 270°\}$. Given an input image x, we randomly apply one of the rotation operations, divide the image into patches, and encode each patch with g_{enc}. A 4-way classifier with convolutional layers g_{rot} is learned on the grid of encoded features to predict the degree of rotation using the cross-entropy loss:

$$\begin{aligned} \mathcal{L}_{Rot} &= -\frac{1}{N}\sum_{x \in \boldsymbol{X}} \log p(\theta | \mathcal{R}_\theta(x)) \\ &= -\frac{1}{N}\sum_{x \in \boldsymbol{X}} \log p(\boldsymbol{y}(\{\boldsymbol{z}\}_{i\times j}) | \mathcal{R}_\theta(x)) \end{aligned} \tag{3}$$

where $\boldsymbol{y}(\{\boldsymbol{z}\}_{i\times j}) = \text{softmax}(g_{\text{rot}}(\{\boldsymbol{z}\}_{i\times j}))$ (i.e. the softmax output of the classifier g_{rot} given $\{\boldsymbol{z}\}_{i\times j}$) and $\{\boldsymbol{z}\}_{i\times j}$ denotes the grid of encoded features $\boldsymbol{z}_{i,j} = g_{\text{enc}}(\mathcal{R}_\theta(x)_{i,j})$.

This loss is jointly optimized with the CPC objective; thus the final loss we optimize for our proposed method is defined as follows:

$$\mathcal{L}_{RotCPC} = \mathcal{L}_{CPC} + \lambda \cdot \mathcal{L}_{Rot} \tag{4}$$

where $\lambda > 0$ is a hyperparameter that balances the contribution of the loss term.

Several preliminary experiments were conducted to determine this formulation, see Appendix for details. The intuition behind Rotated CPC is that since the model has to solve the CPC task on rotated images, it is forced to learn separate latent subspaces for each rotation. This not only allows the model to learn geometrical structures unique to the normal images, but also provides additional robustness in computing the novelty score as we ensemble the scores from each subspace.

3.4 Novelty Score Function

We now define our novelty score function $ns(x)$ used for detecting abnormal samples during inference. Given a test set $\mathcal{S}$, the novelty score of input $x \in \mathcal{S}$ is computed by normalizing the log-likelihood terms introduced in equation 2 and equation 3 with regards to their maximum and minimum values in $\mathcal{S}$, and summing the normalized values as a single score:

$$ns(x) = \textit{norm}_{\mathcal{S}}(\sum_{i,j,k} \log p(\bar{\boldsymbol{z}}_{i+k,j} | \hat{\boldsymbol{z}}_{i+k,j}, \{\bar{\boldsymbol{z}}_l\})) + \textit{norm}_{\mathcal{S}}(\log p(\boldsymbol{y}(\{\boldsymbol{z}\}_{i\times j}) | \mathcal{R}_\theta(x))) \tag{5}$$

where $\textit{norm}_{\mathcal{S}}(L_i) = \frac{L_i - \max_{j\in S} L_j}{\max_{j\in S} L_j - \min_{j\in S} L_j}$.

We can further improve the proposed novelty score by averaging their values across all subspaces associated with the rotations. Specifically, we augment the input x using all rotations in $\mathcal{R}_\theta$ and compute the mean of the scores computed for each rotated x:

$$ns^*(x) = \mathbb{E}_\theta[ns(\mathcal{R}_\theta(x))] \tag{6}$$

4 Experiments

In this section, we describe our experimental setup including datasets, evaluation protocol, and the baseline methods we compare with. We also present quantitative results that show the effectiveness of our Rotated CPC and its corresponding novelty score function. Finally, we provide additional analysis of our proposed method through various ablation experiments. The implementation details can be found in the Appendix.

4.1 Datasets and Evaluation Protocols

We run experiments on four publicly available datasets: MNIST (LeCun, 1998), FashionMNIST (Xiao et al., 2017), CIFAR-10 (Krizhevsky et al., 2009), and COIL-100 (Nene et al., 1996), with image sizes ranging from 28×28 to 640×480 and between $7,200$ and $70,000$ samples. Full details about the datasets can be found in the Appendix.

To generate train-test splits, there are two protocols that are commonly used in the novelty detection problem (Pidhorskyi et al., 2018; Perera et al., 2019):

- Protocol 1: The original training and testing splits of the dataset are combined. 80% of the normal class samples from the combined pool are used as training set whereas the remaining 20% goes to test set. The remaining portion of the test data is randomly sampled from the anomolous classes.
- Protocol 2: The original training and testing splits of the dataset are maintained. The normal set is generated by using only the normal samples from the training split while evaluation is performed on the entire test set.

For MNIST, Fashion-MNIST, and CIFAR-10, we follow protocol 2 and do a one-vs-all evaluation where one class is considered as normal while the other remaining classes are considered anomalous. For COIL-100, we follow protocol 1 and do a multi-class evaluation: we randomly sample p classes as normal classes where $p \in \{1, 4, 7\}$, and the remaining classes are assumed as the abnormal classes.

4.2 BASELINE METHODS

4.2.1 RECONSTRUCTION-BASED METHODS

Deep Structured Energy Based Model (DSEBM) is a model that learns the energy function associated with normal data, and it is trained in a similar way how a denoising autoencoder is trained using score matching (Zhai et al., 2016). **Deep Autoencoding Gaussian Mixture Model (DAGMM)** is a method that optimizes a deep autoencoder to learn latents on which a Gaussian mixture model is jointly fitted (Zong et al., 2018). The reconstruction error and energy criteria are combined to compute the novelty score. Similarly, **Generative Probabilistic Novelty Decection (GPND)** trains an adversarial autoencoder to learn the manifold of the normal data on which a generalized Gaussian distribution is fitted (Pidhorskyi et al., 2018). **Memory-augmented autoencoder (MemAE)** is an autoencoder-based method that utilizes memory banks to learn prototypical features essential to reconstructing normal images (Gong et al., 2019). **Latent Space Autoregression (LSA)** is another autoencoder model that is equipped with an autoregressive density estimator that learns the probability distribution of its latents (Abati et al., 2019). **Adversarially Robust trained Autoencoder (ARAE)** is a model that utilizes adversarial perturbations to promote robust feature learning helpful for novelty detection. **AnoGAN** and **OCGAN** rely on GANs to reconstruct images (Schlegl et al., 2017; Perera et al., 2019). For computing novelty scores, AnoGAN finds a latent that reconstructs the query image and computes the L1 residuals between the query and reconstructed images, whereas OCGAN uses L2 error and/or discriminator score given the reconstructed images.

4.2.2 NON-RECONTRUCTION-BASED METHODS

One-Class Support Vector Machine (OC-SVM) is a Gaussian kernel-based method for one-class classification (Schölkopf et al., 2000). OC-SVM finds a maximum margin hyperplane in feature space that seperates the normal data from the origin. This translates to finding a space in which most normal data exist, and data point outside of this space is considered an outlier. Closely related, but more recent work is the **Deep Support Vector Data Description (Deep SVDD)** (Ruff et al., 2018). Deep SVDD is also a one-class classification method that learns a neural network which maps the normal data into a hypersphere of minimum volume, and the data point outside of this hypersphere is detected as an anomaly. **R-graph** constructs a directed graph based on data representations and uses random walks on the graph to detect outliers (You et al., 2017). A **geometric transformation-based model**, which we denote as **Geo**, is one of the state-of-the-art methods for novelty detection (Golan & El-Yaniv, 2018). The main idea is to train a classifier that discriminates what geometric transformations are applied on the normal images, and use the classifier's softmax activation statistics to detect outliers given transformed test images. **Rot + Trans** is a model based on a pretext task of predicting the rotations as well as horizontal/vertical translations of the image (Hendrycks et al., 2019). **GOAD** is a method that combines one-class classification and transformation-based classification in a unified framework (Bergman & Hoshen, 2020).

4.3 RESULTS

First, we report our model's performance on one-vs-all novelty detection scheme in Table 1, Table 2, and Table 3. Results for all the metrics are copied from each corresponding paper. As can be seen from the tables, our proposed model is able to outperform strong baselines on a majority of image classes, achieving the highest average scores consistently across datasets. In particular, our method is able to perform well for some of the more difficult classes wherein the previous methods struggle. For instance in the MNIST experiment, the digit that many baselines perform the worst is 8; however, Rotated CPC is able to achieve improved performance with a significant margin ($+3.4\%$ absolute from the second best). Additionally in the case of CIFAR-10, Rotated CPC is the only model that exceeds 90% while all other baselines are below 80% for class "airplane."

We notice that the non-reconstruction-based methods in general perform better than reconstruction-based methods as the dataset becomes more complex (i.e. from digits to objects, from grayscale to color, from clean backgrounds to cluttered backgrounds). This gives evidence for the importance of having a novelty detection system that has a more semantic understanding of normality that goes beyond simple reconstruction of the input.

Method	0	1	2	3	4	5	6	7	8	9	Avg
Reconstruction-based											
MemAE	–	–	–	–	–	–	–	–	–	–	97.5
LSA	99.3	**99.9**	95.9	96.6	95.6	96.4	99.4	98.0	95.3	98.1	97.5
ARAE	**99.8**	**99.9**	96.0	97.2	97.0	97.4	**99.5**	96.9	92.4	98.5	97.5
AnoGAN	96.6	99.2	85.0	88.7	89.4	88.3	94.7	93.5	84.9	92.4	91.3
OCGAN	**99.8**	**99.9**	94.2	96.3	97.5	98.0	99.1	**98.1**	93.9	98.1	97.5
Non-Reconstruction-based											
OCSVM	98.8	**99.9**	90.2	95.0	95.5	96.8	97.8	96.5	85.3	95.5	95.1
Deep SVDD	98.6	99.7	91.7	91.9	94.9	88.5	98.3	94.6	93.9	96.5	94.8
Rotated CPC	99.7	99.3	**98.4**	**97.9**	**97.9**	**98.6**	99.1	97.4	**98.7**	**99.0**	**98.6**

Table 1: AUROC values for MNIST dataset. The best performing method in each column is in bold.

Method	0	1	2	3	4	5	6	7	8	9	Avg
Reconstruction-based											
DSEBM	86.0	97.1	85.2	87.3	88.3	87.1	73.4	98.1	71.8	91.6	86.6
DAGMM	71.8	34.0	26.9	57.0	50.4	70.5	48.3	83.5	55.1	42.1	51.8
ARAE	93.7	99.1	91.1	**94.4**	92.3	91.4	83.6	**98.9**	93.9	97.9	93.6
Non-Reconstruction-based											
Deep SVDD	91.9	99.0	89.4	94.2	90.7	91.8	83.4	98.8	90.3	98.2	92.8
Geo	90.8	99.2	92.1	89.9	91.1	93.4	83.3	**98.9**	97.6	**99.4**	93.5
GOAD	**98.9**	99.2	91.4	91.6	90.8	94.8	83.4	97.9	98.5	94.1	94.1
Rotated CPC	93.9	**99.5**	**93.6**	92.5	**94.5**	**95.5**	**85.5**	97.7	**98.8**	99.3	**95.1**

Table 2: AUROC values for Fashion-MNIST dataset. The best performing method in each column is in bold. For the interest of space, we omit class names in the table and list the names here instead: {T-shirt/top, trouser, pullover, dress, coat, sandal, shirt, sneaker, bag, ankle boot}

We also demonstrate that our solution works well in the multi-class setting where more than one class can constitute the normal set. In Table 4, p denotes the number of classes constituting the normal dataset and the percentage denotes the proportion of anomalies in the test set. In every scenario, our proposed metohd is able to outperform state-of-the-art methods in both AUROC and F_1 metrics.

4.4 Ablations

Figure 2 ablates Rotated CPC on multiple components to better understand the improvements provided by the model design choices. We compare the following:

i CPC T: Vanilla CPCv2 with only top-down supervision

ii CPC TBRL: Vanilla CPCv2 with 4 directional supervision

iii Rot CPC T w/o Ens: CPC T with rotation, but without rotation ensembling

iv Rot CPC w/o Ens: CPC TBRL with rotation, but without rotation ensembling

v Rot CPC T: CPC T with rotation and rotation ensembling

vi Rot CPC: CPC TBRL with rotation and rotation ensembling (our full method)

We conduct the ablation experiments on the CIFAR-10 dataset. We are interested in understanding the effects of providing additional supervision, introducing rotation subspaces, and ensembling scores. We notice that the additional supervision introduced by predicting future latents in multiple directions consistently helps performance (i vs. ii , iii vs. iv , v vs. vi). This seems to indicate that techniques that improve the general quality of representations also benefit the downstream novelty detection task. We also observe the effectiveness of introducing rotation (i vs. iii , ii vs. iv). The performance gain from ii to iv is especially interesting. Rotation and 4 directional supervision could be redundant since the latter is equivalent to rotating the features and doing top-down prediction.

Method	0	1	2	3	4	5	6	7	8	9	Avg
Reconstruction-based											
DSEBM	56.0	48.3	61.9	50.1	73.3	60.5	68.4	53.3	73.9	63.6	60.9
DAGMM	41.4	57.1	53.8	51.2	52.2	49.3	64.9	55.3	51.9	54.2	53.1
MemAE	–	–	–	–	–	–	–	–	–	–	60.1
LSA	73.5	58.0	69.0	54.2	76.1	54.6	75.1	53.5	71.7	54.8	64.1
AnoGAN	67.1	54.7	52.9	54.5	65.1	60.3	58.5	62.5	75.8	66.5	61.8
OCCAN	75.7	53.1	64.0	62.0	72.3	62.0	72.3	57.5	82.0	55.4	65.7
Non-Reconstruction-based											
OCSVM	63.0	44.0	64.9	48.7	73.5	50.0	72.5	53.3	64.9	50.8	58.6
Deep SVDD	61.7	65.9	50.8	59.1	60.9	65.7	67.7	67.3	75.9	73.1	64.8
Geo	74.7	95.7	78.1	72.4	87.8	87.8	83.4	95.5	93.3	91.3	86.0
GOAD	77.2	96.7	83.3	77.7	87.8	87.8	90.0	96.1	93.8	92.0	88.2
Rot + Trans	77.5	96.9	**87.3**	**80.9**	**92.7**	**90.2**	**90.9**	**96.5**	95.2	93.3	90.1
Rotated CPC	**91.4**	**98.0**	84.8	80.0	86.1	88.4	**90.9**	95.6	**96.0**	**93.9**	**90.5**

Table 3: AUROC values for CIFAR-10 dataset. The best performing method in each column is in bold. For the interest of space, we omit class names in the table and list the names here instead: {airplane, automobile, bird, cat, deer, dog, frog, horse, ship, truck}

Yet, the performance boost indicates that there is some additional benefits to be gained from rotating images. We speculate that since convolutional neural networks are not rotation-invariant by design, each rotation gets considered as a different domain of normality which the CPC module needs to deal with. Finally, we show that ensembling the scores of different rotations during inference is beneficial (iii vs. v , iv vs. vi).

	AUC	F_1
$p = 1$, Anomalous: 50%		
R-graph	0.997	0.990
GPND	0.968	0.979
ARAE	0.998	0.993
Rotated CPC	**1.000**	**0.999**
$p = 4$, Anomalous: 25%		
R-graph	0.996	0.970
GPND	0.945	0.960
ARAE	0.997	0.973
Rotated CPC	**1.000**	**0.999**
$p = 7$, Anomalous: 15%		
R-graph	0.996	0.955
GPND	0.919	0.941
ARAE	0.993	0.941
Rotated CPC	**1.000**	**0.999**

Table 4: AUC and F_1 scores for COIL-100 dataset.

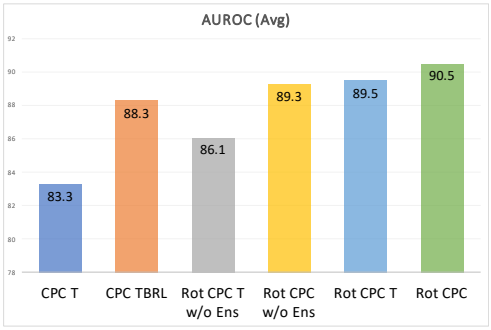

Figure 2: Ablations on CIFAR-10.

5 Conclusion

Significant advancement has been made recently in contrastive learning as a powerful framework for visual representation learning. At the same time, impressive progress has been observed in novelty detection where conventional reconstruction-based methods are being overshadowed by methods that alleviate or even remove the need for reconstruction. These two advancements naturally lend themselves to the idea of leveraging contrastive learning to solve novelty detection problem. In this paper, we explore this idea and demonstrate that Contrastive Predictive Coding can be used to achieve strong results on multiple benchmark datasets. We additionally show that a simple augmentation, where the CPC objective is optimized in rotation subspaces, can further help improve the performance. Our results open up new avenues for future novelty detection research as more advancements in contrastive methods take place.

REFERENCES

Davide Abati, Angelo Porrello, Simone Calderara, and Rita Cucchiara. Latent space autoregression for novelty detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 481–490, 2019.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. In *Advances in Neural Information Processing Systems*, pp. 15535–15545, 2019.

Liron Bergman and Yedid Hoshen. Classification-based anomaly detection for general data. *arXiv preprint arXiv:2005.02359*, 2020.

Gilles Blanchard, Gyemin Lee, and Clayton Scott. Semi-supervised novelty detection. *The Journal of Machine Learning Research*, 11:2973–3009, 2010.

Raghavendra Chalapathy and Sanjay Chawla. Deep learning for anomaly detection: A survey. *arXiv preprint arXiv:1901.03407*, 2019.

Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):1–58, 2009.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020.

David A Clifton, Peter R Bannister, and Lionel Tarassenko. A framework for novelty detection in jet engine vibration data. In *Key engineering materials*, volume 347, pp. 305–310. Trans Tech Publ, 2007.

Carl Doersch, Abhinav Gupta, and Alexei A. Efros. Unsupervised visual representation learning by context prediction. In *International Conference on Computer Vision*, 2015.

Pedro Garcia-Teodoro, Jesus Diaz-Verdejo, Gabriel Maciá-Fernández, and Enrique Vázquez. Anomaly-based network intrusion detection: Techniques, systems and challenges. *computers & security*, 28(1-2):18–28, 2009.

Izhak Golan and Ran El-Yaniv. Deep anomaly detection using geometric transformations. In *Advances in Neural Information Processing Systems*, pp. 9758–9769, 2018.

Dong Gong, Lingqiao Liu, Vuong Le, Budhaditya Saha, Moussa Reda Mansour, Svetha Venkatesh, and Anton van den Hengel. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1705–1714, 2019.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*, 2020.

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

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*, pp. 770–778, 2016.

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*, pp. 9729–9738, 2020.

Olivier J Hénaff, Aravind Srinivas, Jeffrey De Fauw, Ali Razavi, Carl Doersch, SM Eslami, and Aaron van den Oord. Data-efficient image recognition with contrastive predictive coding. *arXiv preprint arXiv:1905.09272*, 2019.

Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. *arXiv preprint arXiv:1812.04606*, 2018.

Dan Hendrycks, Mantas Mazeika, Saurav Kadavath, and Dawn Song. Using self-supervised learning can improve model robustness and uncertainty. In *Advances in Neural Information Processing Systems*, pp. 15663–15674, 2019.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

Rafal Jozefowicz, Oriol Vinyals, Mike Schuster, Noam Shazeer, and Yonghui Wu. Exploring the limits of language modeling. *arXiv preprint arXiv:1602.02410*, 2016.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Yann LeCun. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998.

Andriy Mnih and Yee Whye Teh. A fast and simple algorithm for training neural probabilistic language models. *arXiv preprint arXiv:1206.6426*, 2012.

Sameer A Nene, Shree K Nayar, Hiroshi Murase, et al. Columbia object image library (coil-100). 1996.

Aaron van den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. *arXiv preprint arXiv:1601.06759*, 2016.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018a.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018b.

Pramuditha Perera, Ramesh Nallapati, and Bing Xiang. Ocgan: One-class novelty detection using gans with constrained latent representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2898–2906, 2019.

Stanislav Pidhorskyi, Ranya Almohsen, and Gianfranco Doretto. Generative probabilistic novelty detection with adversarial autoencoders. In *Advances in neural information processing systems*, pp. 6822–6833, 2018.

Lukas Ruff, Robert Vandermeulen, Nico Goernitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. Deep one-class classification. In *International conference on machine learning*, pp. 4393–4402, 2018.

Mohammadreza Salehi, Atrin Arya, Barbod Pajoum, Mohammad Otoofi, Amirreza Shaeiri, Mohammad Hossein Rohban, and Hamid R Rabiee. Arae: Adversarially robust training of autoencoders improves novelty detection. *arXiv preprint arXiv:2003.05669*, 2020.

Thomas Schlegl, Philipp Seeböck, Sebastian M Waldstein, Ursula Schmidt-Erfurth, and Georg Langs. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In *International conference on information processing in medical imaging*, pp. 146–157. Springer, 2017.

Bernhard Schölkopf, Robert C Williamson, Alex J Smola, John Shawe-Taylor, and John C Platt. Support vector method for novelty detection. In *Advances in neural information processing systems*, pp. 582–588, 2000.

Yan Xia, Xudong Cao, Fang Wen, Gang Hua, and Jian Sun. Learning discriminative reconstructions for unsupervised outlier removal. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1511–1519, 2015.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Chong You, Daniel P Robinson, and René Vidal. Provable self-representation based outlier detection in a union of subspaces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3395–3404, 2017.

Shuangfei Zhai, Yu Cheng, Weining Lu, and Zhongfei Zhang. Deep structured energy based models for anomaly detection. *arXiv preprint arXiv:1605.07717*, 2016.

Arthur Zimek, Erich Schubert, and Hans-Peter Kriegel. A survey on unsupervised outlier detection in high-dimensional numerical data. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 5(5):363–387, 2012.

Bo Zong, Qi Song, Martin Renqiang Min, Wei Cheng, Cristian Lumezanu, Daeki Cho, and Haifeng Chen. Deep autoencoding gaussian mixture model for unsupervised anomaly detection. In *International Conference on Learning Representations*, 2018.

A APPENDIX

A.1 ADDITIONAL DETAILS ON DATASETS

Dataset	Image Size	Classes	Train	Test	Total
MNIST (LeCun, 1998)	$1 \times 28 \times 28$	10	60,000	10,000	70,000
FashionMNIST (Xiao et al., 2017)	$1 \times 28 \times 28$	10	60,000	10,000	70,000
CIFAR-10 (Krizhevsky et al., 2009)	$3 \times 32 \times 32$	10	50,000	10,000	60,000
COIL-100 (Nene et al., 1996)	$3 \times 640 \times 480$	100	–	–	7,200

Table 5: Utilized datasets summary

A.2 IMPLEMENTATION DETAILS

For all datasets, we resize the image to 128×128 and extract patches of size 32×32 where 16 pixels overlap between patches. This results in a 7×7 grid of patches for a single image. During training, each patch is randomly cropped by 28×28 and zero-padded to recover the original size. For color images, stochastic color dropping is applied to the patch to avoid learning degenerate solutions based on chromatic aberration. The patch encoder g_{enc} is implemented using ResNet-18 (He et al., 2016), but with BatchNorm replaced with LayerNorm. The rotation classifier g_{rot} is a model with one convolutional layer with ReLU activation, followed by global average pooling and a linear layer for prediction. The length of future prediction steps $k \in \{2, 3\}$. The entire model is learned by optimizing the loss $\mathcal{L}_{RotCPC}$ (Equation 4) where $\lambda = 1$ for all experiments. We use ADAM optimizer (Kingma & Ba, 2014) with a learning rate of 0.0002. For MNIST, the model is trained for 50 epochs while for the other datasets we train for 500 epochs. The batch size is 32 for all experiments.

A.3 PRELIMINARY EXPERIMENTS REGARDING THE FORMULATION OF ROTATED CPC

Several experiments were conducted prior to arriving at the proposed formulation of Rotated CPC. Before deciding to augment CPC with rotation, we first experimented with adding a decoder on top of the CPC that reconstructs input image given encoded patch features from g_{enc}. The decoder was jointly trained with CPC using $L2$ loss. As can be seen from Table 6, augmenting with reconstruction task only led to a small increase in performance compared to augmenting with image-level rotation. Also, we have tried patch-level rotation; however, we observed a regression in performance. We noticed that the CPC task becomes too difficult to solve when patches are rotated independently, and as a result the model started converging to a bad local minimum.

Method	AUROC (Avg)
CPC TBRL	88.3
CPC TBRL + Reconstruction	88.9
CPC TBRL + Patch-level Rotation	86.1
CPC TBRL + Image-level Rotation	89.3

Table 6: Preliminary experiments on CIFAR-10.