Anomaly Detection via Reverse Distillation from One-Class Embedding

Hanqiu Deng Xingyu Li
Department of Electrical and Computer Engineering, University of Alberta
{hanqiu1,xingyu}@ualberta.ca

Abstract

Knowledge distillation (KD) achieves promising results on the challenging problem of unsupervised anomaly detection (AD). The representation discrepancy of anomalies in the teacher-student (T-S) model provides essential evidence for AD. However, using similar or identical architectures to build the teacher and student models in previous studies hinders the diversity of anomalous representations. To tackle this problem, we propose a novel T-S model consisting of a teacher encoder and a student decoder and introduce a simple yet effective "reverse distillation" paradigm accordingly. Instead of receiving raw images directly, the student network takes teacher model’s one-class embedding as input and targets to restore the teacher’s multi-scale representations. Inherently, knowledge distillation in this study starts from abstract, high-level presentations to low-level features. In addition, we introduce a trainable one-class bottleneck embedding (OCBE) module in our T-S model. The obtained compact embedding effectively preserves essential information on normal patterns, but abandons anomaly perturbations. Extensive experimentation on AD and one-class novelty detection benchmarks shows that our method surpasses SOTA performance, demonstrating our proposed approach’s effectiveness and generalizability.

1. Introduction

Anomaly detection (AD) refers to identifying and localizing anomalies with limited, even no, prior knowledge of abnormality. The wide applications of AD, such as industrial defect detection [3], medical out-of-distribution detection [50], and video surveillance [24], makes it a critical task as well as a spotlight. In the context of unsupervised AD, no prior information on anomalies is available. Instead, a set of normal samples is provided for reference. To tackle this problem, previous efforts attempt to construct various self-supervision tasks on those anomaly-free samples. These tasks include, but not limited to, sample reconstruction [2,5,11,16,26,34,38,48], pseudo-outlier augmentation [23,42,46], knowledge distillation [4,33,39], etc.

In this study, we tackle the problem of unsupervised anomaly detection from the knowledge distillation-based point of view. In knowledge distillation (KD) [6,15], knowledge is transferred within a teacher-student (T-S) pair. In the context of unsupervised AD, since the student experiences only normal samples during training, it is likely to generate discrepant representations from the teacher when a query is anomalous. This hypothesis forms the basis of KD-based methods for anomaly detection. However, this hypothesis is not always true in practice due to (1) the identical or similar architectures of the teacher and student networks (i.e., non-distinguishing filters [33]) and (2) the same data flow in the T-S model during knowledge transfer/distillation. Though the use of a smaller student network partially addresses this issue [33,39], the weaker representation capability of shallow architectures hinders the model from precisely detecting and localizing anomalies.

To holistically address the issue mentioned above, we propose a new paradigm of knowledge distillation, namely Reverse Distillation, for anomaly detection. We use simple diagrams in Fig. 2 to highlight the systematic difference between conventional knowledge distillation and the proposed reverse distillation. First, unlike the conventional knowledge distillation framework where both teacher and student adopt the encoder structure, the T-S model in our
reverse distillation consists of heterogeneous architectures: a teacher encoder and a student decoder. Second, instead of directly feeding the raw data to the T-S model simultaneously, the student decoder takes the low-dimensional embedding as input, targeting to mimic the teacher’s behavior by restoring the teacher model’s representations in different scales. From the regression perspective, our reverse distillation uses the student network to predict the representation of the teacher model. Therefore, “reverse” here indicates both the reverse shapes of teacher encoder and student decoder and the distinct knowledge distillation order where high-level representation is first distilled, followed by low-level features. It is noteworthy that our reverse distillation presents two significant advantages: i) Non-similarity structure. In the proposed T-S model, one can consider the teacher encoder as a down-sampling filter and the student decoder as an up-sampling filter. The “reverse structures” avoid the confusion caused by non-distinguishing filters [33] as we discussed above. ii) Compactness embedding. The low-dimensional embedding fed to the student decoder acts as an information bottleneck for normal pattern restoration. Let’s formulate anomaly features as perturbations on normal patterns. Then the compact embedding helps to prohibit the propagation of such unusual perturbations to the student model and thus boosts the T-S model’s representation discrepancy on anomalies. Notably, traditional AE-based methods [5, 11, 16, 26] detect anomalies utilising pixel differences, whereas we perform discrimination with dense descriptive features. Deep features as region-aware descriptors provide more effective discriminative information than per-pixel in images.

In addition, since the compactness of the bottleneck embedding is vital for anomaly detection (as discussed above), we introduce a one-class bottleneck embedding (OCBE) module to condense the feature codes further. Our OCBE module consists of a multi-scale feature fusion (MFF) block and one-class embedding (OCE) block, both jointly optimized with the student decoder. Notably, the former aggregates low- and high-level features to construct a rich embedding for normal pattern reconstruction. The latter targets to retain essential information favorable for the student to decode out the teacher’s response.

We perform extensive experiments on public benchmarks. The experimental results indicate that our reverse distillation paradigm achieves comparable performance with prior arts. The proposed OCBE module further improves the performance to a new state-of-the-art (SOTA) record. Our main contributions are summarized as follows:

- We introduce a simple, yet effective Reverse Distillation paradigm for anomaly detection. The encoder-decoder structure and reverse knowledge distillation strategy holistically address the non-distinguishing filter problem in conventional KD models, boosting the T-S model’s discrimination capability on anomalies.
- We propose a one-class bottleneck embedding module to project the teacher’s high-dimensional features to a compact one-class embedding space. This innovation facilitates retaining rich yet compact codes for anomaly-free representation restoration at the student.
- We perform extensive experiments and show that our approach achieves new SOTA performance.

2. Related Work

This section briefly reviews previous efforts on unsupervised anomaly detection. We will highlight the similarity and difference between the proposed method and prior arts.

Classical anomaly detection methods focus on defining a compact closed one-class distribution using normal support vectors. The pioneer studies include one-class support vector machine (OC-SVM) [35] and support vector data description (SVDD) [36]. To cope with high-dimensional data, DeepSVDD [31] and PatchSVDD [43] estimate data representations through deep networks.

Another unsupervised AD prototype is the use of generative models, such as AutoEncoder (AE) [19] and Generative Adversarial Nets (GAN) [12], for sample reconstruction. These methods rely on the hypothesis that generative models trained on normal samples only can successfully reconstruct anomaly-free regions, but fail for anomalous regions [2, 5, 34]. However, recent studies show that deep models generalize so well that even anomalous regions can be well-restored [46]. To address this issue, memory mechanism [11, 16, 26], image masking strategy [42, 46] and pseudo-anomaly [28, 45] are incorporated in reconstruction-based methods. However, these methods still lack a strong discriminating ability for real-world anomaly detection [3, 5]. Recently, Metaformer (MF) [40] proposes the use of meta-learning [9] to bridge model adaptation and reconstruction gap for reconstruction-based approaches. Notably, the proposed reverse knowledge distillation also adopts the encoder-decoder architecture, but it
differs from construction-based methods in two-folds. First, the encoder in a generative model is jointly trained with the decoder, while our reverse distillation freezes a pre-trained model as the teacher. Second, instead of pixel-level reconstruction error, it performs anomaly detection on the semantic feature space.

Data augmentation strategy is also widely used. By adding pseudo anomalies in the provided anomaly-free samples, the unsupervised task is converted to a supervised learning task [23, 42, 46]. However, these approaches are prone to bias towards pseudo outliers and fail to detect a large variety of anomaly types. For example, CutPaste [23] generates pseudo outliers by adding small patches onto normal images and trains a model to detect these anomalous regions. Since the model focuses on detecting local features such as edge discontinuity and texture perturbations, it fails to detect and localize large defects and global structural anomalies as shown in Fig. 6.

Recently, networks pre-trained on the large dataset are proven to be capable of extracting discriminative features for anomaly detection [7, 8, 23, 25, 29, 30]. With a pre-trained model, memorizing its anomaly-free features helps to identify anomalous samples [7, 29]. The studies in [8, 30] show that using the Mahalanobis distance to measure the similarity between anomalies and anomaly-free features leads to accurate anomaly detection. Since these methods require memorizing all features from training samples, they are computationally expensive.

Knowledge distillation from pre-trained models is another potential solution to anomaly detection. In the context of unsupervised AD, since the student model is exposed to anomaly-free samples in knowledge distillation, the T-S model is expected to generate discrepant features on anomalies in inference [4, 33, 39]. To further increase the discriminating capability of the T-S model on various types of abnormalities, different strategies are introduced. For instance, in order to capture multi-scale anomaly, US [4] ensembles several models trained on normal data at different scales, and MKD [33] propose to use multi-level features alignment. It should be noted that though the proposed method is also based on knowledge distillation, our reverse distillation is the first to adopt an encoder and a decoder to construct the T-S model. The heterogeneity of the teacher and student networks and reverse data flow in knowledge distillation distinguishes our method from prior arts.

3. Our Approach

Problem formulation: Let \( T^t = \{I_1^t, \ldots, I_n^t\} \) be a set of available anomaly-free images and \( T^q = \{I_1^q, \ldots, I_m^q\} \) be a query set containing both normal and abnormal samples. The goal is to train a model to recognize and localize anomalies in the query set. In the anomaly detection setting, normal samples in both \( T^t \) and \( T^q \) follow the same distribution. Out-of-distribution samples are considered anomalies.

System overview: Fig. 3 depicts the proposed reverse distillation framework for anomaly detection. Our reverse distillation framework consists of three modules: a fixed pre-trained teacher encoder \( E \), a trainable one-class bottleneck embedding module, and a student decoder \( D \). Given an input sample \( I \in T^q \), the teacher \( E \) extracts multi-scale representations. We propose to train a student \( D \) to restore the features from the bottleneck embedding. During testing/inference, the representation extracted by the teacher \( E \) can capture abnormal, out-of-distribution features in anomalous samples. However, the student decoder \( D \) fails to reconstruct these anomalous features from the corresponding embedding. The low similarity of anomalous representations in the proposed T-S model indicates a high abnormality score. We argue that the heterogeneous encoder and decoder structures and reverse knowledge distillation order contribute a lot to the discrepant representations of anomalies. In addition, the trainable OCBE module further condenses the multi-scale patterns into an extremely low-dimensional space for downstream normal representation reconstruction. This further improves feature discrepancy on anomalies in our T-S model, as abnormal representations generated by the teacher model are likely to be abandoned by OCBE. In the rest of this section, we first specify the reverse distillation paradigm. Then, we elaborate on the OCBE module. Finally, we describe anomaly detection and localization using reserve distillation.

3.1. Reverse Distillation

In conventional KD, the student network adopts a similar or identical neural network to the teacher model, accepts raw data/images as input, and targets to match its feature activations to the teacher’s [4, 33]. In the context of one-class distillation for unsupervised AD, the student model is expected to generate highly different representations from the teacher when the queries are anomalous samples [11, 26]. However, the activation discrepancy on anomalies vanishes sometimes, leading to anomaly detection failure. We argue that this issue is attributed to the similar architectures of the teacher and student nets and the same data flow during T-S knowledge transfer. To improve the T-S model’s representation diversity on unknown, out-of-distribution samples, we propose a novel reverse distillation paradigm, where the T-S model adopts the encoder-decoder architecture and knowledge is distilled from teacher’s deep layers to its early layers, i.e., high-level, semantic knowledge being transferred to the student first. To further facilitate the one-class distillation, we designed a trainable OCEB module to connect the teacher and student models (Sec. 3.2).

In the reverse distillation paradigm, the teacher encoder \( E \) aims to extract comprehensive representations. We follow previous work and use a pre-trained encoder on Image

\( I \) follows the same distribution. The probability that \( I \) is normal can be computed as:

\[
\pi(q | \mathcal{I}) = \frac{1}{1 + e^{-\frac{\text{Mahalanobis}(q, \mathcal{I})}{\sigma}}} \quad \text{for} \quad q \in \mathcal{I}
\]
Figure 3. Overview of our reverse distillation framework for anomaly detection and localization. (a) Our model consists of a pre-trained teacher encoder $E$, a trainable one-class bottleneck embedding module (OCBE), and a student decoder $D$. We use a multi-scale feature fusion (MFF) block to ensemble low- and high-level features from $E$ and map them onto a compact code by one-class embedding (OCE) block. During training, the student $D$ learns to mimic the behavior of $E$ by minimizing the similarity loss $L$. (b) During inference, $E$ extracts the features truthfully, while $D$ outputs anomaly-free ones. A low similarity between the feature vectors at the corresponding position of $E$ and $D$ implies an abnormality. (c) The final prediction is calculated by the accumulation of multi-scale similarity maps $M$.

gNet [21] as our backbone $E$. To avoid the T-S model converging to trivial solutions, all parameters of teacher $E$ are frozen during knowledge distillation. We show in the ablation study that both ResNet [14] and WideResNet [44] are good candidates, as they are capable of extracting rich features from images [4, 8, 23, 29].

To match the intermediate representations of $E$, the architecture of student decoder $D$ is symmetrical but reversed compared to $E$. The reverse design facilitates eliminating the response of the student network to abnormalities, while the symmetry allows it to have the same representation dimension as the teacher network. For instance, when we take ResNet as the teacher model, the student $D$ consists of several residual-like decoding blocks for mirror symmetry. Specifically, the downsampling in ResNet is realized by a convolutional layer with a kernel size of 1 and a stride of 2 [14]. The corresponding decoding block in the student $D$ adopts deconvolutional layers [47] with a kernel size of 2 and a stride of 2. More details on the student decoder design are given in Supplementary Material.

In our reverse distillation, the student decoder $D$ targets to mimic the behavior of the teacher encoder $E$ during training. In this work, we explore multi-scale feature-based distillation for anomaly detection. The motivation behind this is that shallow layers of a neural network extract local descriptors for low-level information (e.g., color, edge, texture, etc.), while deep layers have wider receptive fields, capable of characterizing regional/global semantic and structural information. That is, low similarity of low- and high-level features in the T-S model suggests local abnormalities and regional/global structural outliers, respectively.

Mathematically, let $\phi$ indicates the projection from raw data $I$ to the one-class bottleneck embedding space, the paired activation correspondence in our T-S model is $\{f^E_k = E^k(I), f^D_k = D^k(\phi)\}$, where $E^k$ and $D^k$ represent the $k^{th}$ encoding and decoding block in the teacher and student model, respectively. $f^E_k, f^D_k \in \mathbb{R}^{C_k \times H_k \times W_k}$, where $C_k, H_k$ and $W_k$ denote the number of channels, height and width of the $k^{th}$ layer activation tensor. For knowledge transfer in the T-S model, the cosine similarity is taken as the KD loss, as it is more precisely to capture the relation in both high- and low-dimensional information [37, 49]. Specifically, for feature tensors $f^E_k$ and $f^D_k$, we calculate their vector-wise cosine similarity loss along the channel axis and obtain a 2-D anomaly map $M^k \in \mathbb{R}^{H_k \times W_k}$:

$$M^k(h, w) = 1 - \frac{(f^E_k(h, w))^T \cdot f^D_k(h, w)}{\left\| f^E_k(h, w) \right\| \cdot \left\| f^D_k(h, w) \right\|}.$$  (1)
A large value in $M^k$ indicates high anomaly in that location. Considering the multi-scale knowledge distillation, the scalar loss function for student’s optimization is obtained by accumulating multi-scale anomaly maps:

$$L_{KD} = \sum_{k=1}^{K} \left\{ \frac{1}{H_k W_k} \sum_{h=1}^{W_k} \sum_{w=1}^{H_k} M^k(h, w) \right\},$$

where $K$ indicates the number of feature layers used in the experiment.

### 3.2. One-Class Bottleneck Embedding

Since the student model $D$ attempts to restore representations of a teacher model $E$ in our reverse knowledge distillation paradigm, one can directly feed the activation output of the last encoding block in backbone to $D$. However, this naive connection has two shortfalls. First, the teacher model in KD usually has a high capacity. Though the high-capacity model helps extract rich features, the obtained high-dimensional descriptors likely have a considerable redundancy. The high freedom and redundancy of representations are harmful to the student model to decode the essential anomaly-free features. Second, the activation of the last encoder block in backbone usually characterizes semantic and structural information of the input data. Due to the reverse order of knowledge distillation, directly feeding this high-level representation to the student encoder set a challenge for low-level features reconstruction. Previous efforts on data reconstruction usually introduce skip paths to connect the encoder and decoder. However, this approach doesn’t work in knowledge distillation, as the skip paths leak anomaly information to the student during inference.

To tackle the first shortfall above in one-class distillation, we introduce a trainable one-class embedding block to project the teacher model’s high-dimensional representations into a low-dimensional space. Let’s formulate anomaly features as perturbations on normal patterns. Then the compact embedding acts as an information bottleneck and helps to prohibit the propagation of unusual perturbations to the student model, therefore boosting the T-S model’s representation discrepancy on anomalies. In this study, we adopt the 4th residue block of ResNet [14] as the one-class embedding block.

To address the problem on low-level feature restoration at decoder $D$, the MFF block concatenates multi-scale representations before one-class embedding. To achieve representation alignment in feature concatenation, we down-sample the shallow features through one or more $3 \times 3$ convolutional layers with stride of 2, followed by batch normalization and ReLU activation function. Then a $1 \times 1$ convolutional layer with stride of 1 and a batch normalization with relu activation are exploited for a rich yet compact feature.

We depict the OCBE module in Fig. 4, where MFF aggregates low- and high-level features to construct a rich embedding for normal pattern reconstruction and OCE targets to retain essential information favorable for the student to decode out the teacher’s response. The convolutional layers in grey and ResBlock in green in Fig. 4 are trainable and optimized jointly with the student model $D$ during knowledge distillation on normal samples.

### 3.3. Anomaly Scoring

At the inference stage, we first consider the measurement of pixel-level anomaly score for anomaly localization (AL). When a query sample is anomalous, the teacher model is capable of reflecting abnormality in its features. However, the student model is likely to fail in abnormal feature restoration, since the student decoder only learns to restore anomaly-free representations from the compact one-class embedding in knowledge distillation. In other words, the student $D$ generates discrepant representations from the teacher when the query is anomalous. Following Eq. (1), we obtain a set of anomaly maps from T-S representation pairs, where the value in a map $M_k$ reflects the point-wise anomaly of the $k^{th}$ feature tensors. To localize anomalies in a query image, we up-samples $M^k$ to image size. Let $\Psi$ denotes the bilinear up-sampling operation used in this study. Then a precise score map $S_{AL}$ is formulated as the pixel-wise accumulation of all anomaly maps:

$$S_{AL} = \sum_{i=1}^{L} \Psi(M^i).$$

In order to remove the noises in the score map, we smooth $S_{AL}$ by a Gaussian filter.

For anomaly detection, averaging all values in a score map $S_{AL}$ is unfair for samples with small anomalous regions. The most responsive point exists for any size of
Table 1. Anomaly Detection results on MVTec [3]. For each category with images of 256 × 256 resolution, methods achieved for the top two AUROC (%) are highlighted in bold. Our method ranks first according to the average scores of textures, objects and overall.

5. Experiments and Discussions

Empirical evaluations are carried on both the MVTec anomaly detection and localization benchmark and unsupervised one-class novelty detection datasets. In addition, we perform ablation study on the MVTec benchmark, investigating the effects of different modules/blocks on the final results.

4.1. Anomaly Detection and Localization

Dataset. MVTec [3] contains 15 real-world datasets for anomaly detection, with 5 classes of textures and 10 classes of objects. The training set comprises a total of 3,629 anomaly-free images. The test set has both anomalous and anomaly-free images, totaling 1,725. Each class has multiple defects for testing. In addition, pixel-level annotations are available in the test dataset for anomaly localization.

Experimental settings. All images in MVTec are resized to a specific resolution (e.g. 128 × 128, 256 × 256 etc.). Following convention in prior works, anomaly detection and localization are performed on one category at a time. In this experiment, we adopt WideResNet50 as Backbone E in our T-S model. We also report the AD results with ResNet18 and ResNet50 in ablation study. To train our reserve distillation model, we utilize Adam optimizer [18] with $\beta = (0.5, 0.999)$. The learning rate is set to 0.005. We train 200 epochs with a batch size of 16. A Gaussian filter with $\sigma = 4$ is used to smooth the anomaly score map (as described in Sec. 3.3).

For Anomaly detection, we take area under the receiver operating characteristic (AUROC) as the evaluation metric. We include prior arts in this experiments, including MKD [33], GT [10], GANomaly (GN) [2], Uninformed Student (US) [4], PSVDD [43], DAAD [16], MetaFormer (MF) [40], PaDiM (WResNet50) [8] and CutPaste [23].

For Anomaly Localization, we report both AUROC and per-region-overlap (PRO) [4]. Different from AUROC, which is used for per-pixel measurement, the PRO score treats anomaly regions with any size equally. The comparison baselines includes MKD [33], US [4], MF [40], SPADE (WResNet50) [7,29], PaDiM (WResNet50) [8], RIAD [46] and CutPaste [23].

Experimental results and discussions. Anomaly detection results on MVTec are shown in Tab. 1. The average outcome shows that our method exceeds SOTA by 2.5%. For textures and objects, our model reaches new SOTA of 99.5% and 98.0% of AUROC, respectively. The statistics of the anomaly scores are shown in Fig. 5. The non-overlap distribution of normal (blue) and anomalies (red) indicates the strong AD capability in our T-S model.

Quantitative results on anomaly localization are summarized in Tab. 2. For both AUROC and PRO average scores over all categories, our approach surpasses state-of-the-art with 97.8% and 93.9%. To investigate the robustness of our method to various anomalies, we classify the defect types into two categories: large defects or structural anomalies and tiny or inconspicuous defects, and qualitative evaluate the performance by visualization in Fig. 6 and Fig. 7. Compared to the runner-up (i.e. CutPaste [23]) in Tab. 1, our method produces a significant response to the whole anomalous region. Hence, we define the maximum value in $S_{AL}$ as sample-level anomaly score $S_{AD}$. The intuition is that no significant response exists in their anomaly score map for normal samples.

Objects

| Category | Texture | Image Size 128 | Ours | GT [10] | GN [2] | US [4] | PSVDD [43] | DAAD [16] | MF [40] | PaDiM [8] | CutPaste [23] | Ours |
|----------|---------|----------------|------|---------|--------|-------|-----------|-----------|--------|---------|-------------|------|
| Bottle   | Carpet  | 79.3 | 97.2 | 43.7 | 69.9 | 91.6 | 92.9 | 86.6 | 94.0 | 99.3 | 99.3 | 99.3 |
| Grid     | Grid    | 78.0 | 95.7 | 61.9 | 70.8 | 81.0 | 94.6 | 95.7 | 85.9 | 96.7 | 100 | 100 |
| Leather  | Leather | 95.1 | 100 | 84.1 | 84.2 | 88.2 | 90.9 | 86.2 | 92.2 | 100 | 100 | 100 |
| Tile     | Tile    | 91.6 | 99.4 | 41.7 | 79.4 | 99.1 | 97.8 | 88.2 | 99.0 | 98.1 | 94.6 | 99.3 |
| Wood     | Wood    | 94.3 | 98.8 | 61.1 | 83.4 | 97.7 | 96.5 | 98.2 | 99.2 | 99.2 | 99.1 | 99.2 |
| Average  | Average | 87.7 | 98.6 | 58.5 | 77.5 | 91.3 | 94.5 | 91.0 | 95.5 | 98.8 | 97.5 | 99.5 |

| Category | Objects | Image Size 256 | Ours | GT [10] | GN [2] | US [4] | PSVDD [43] | DAAD [16] | MF [40] | PaDiM [8] | CutPaste [23] | Ours |
|----------|---------|----------------|------|---------|--------|-------|-----------|-----------|--------|---------|-------------|------|
| Bottle   | Bottle  | 99.4 | 100 | 74.4 | 89.2 | 99.0 | 98.6 | 97.6 | 99.1 | 99.9 | 98.2 | 100 |
| Cable    | Cable   | 89.2 | 97.1 | 78.3 | 75.7 | 86.2 | 90.3 | 84.4 | 97.1 | 92.7 | 81.2 | 95.0 |
| Capsule  | Capsule | 80.5 | 89.5 | 67.0 | 73.2 | 86.1 | 76.7 | 76.7 | 87.5 | 91.3 | 98.2 | 96.3 |
| Hazelnut | Hazelnut| 98.4 | 99.8 | 35.9 | 78.5 | 93.1 | 92.0 | 92.1 | 99.4 | 92.0 | 98.3 | 99.9 |
| Metal Nut| Metal Nut| 73.6 | 99.2 | 81.3 | 70.0 | 82.0 | 94.0 | 75.8 | 96.2 | 98.7 | 99.9 | 100 |
| Pill     | Pill    | 82.7 | 93.3 | 63.0 | 74.3 | 87.9 | 86.1 | 90.0 | 90.1 | 93.3 | 94.9 | 96.6 |
| Screw   | Screw   | 83.3 | 91.1 | 50.0 | 74.6 | 54.9 | 81.3 | 98.7 | 97.5 | 85.8 | 88.7 | 97.0 |
| Toothbrush| Toothbrush| 92.2 | 90.3 | 97.2 | 65.3 | 95.3 | 100 | 99.2 | 100 | 96.1 | 99.4 | 99.5 |
| Transistor| Transistor| 85.6 | 99.5 | 86.9 | 79.2 | 81.8 | 91.5 | 87.6 | 94.4 | 97.4 | 96.1 | 96.7 |
| Zipper  | Zipper  | 93.2 | 94.3 | 82.0 | 74.5 | 91.9 | 97.9 | 85.9 | 98.6 | 90.3 | 99.9 | 98.5 |
| Average  | Average | 87.8 | 95.4 | 71.6 | 75.5 | 83.8 | 90.8 | 88.8 | 96.0 | 93.8 | 95.5 | 98.0 |

| Total Average | 87.8 | 96.3 | 67.2 | 76.2 | 87.7 | 92.1 | 89.3 | 93.8 | 93.5 | 96.7 | 98.5 |
Table 2. Anomaly Localization results with AUROC and PRO on MVTec [3]. AUROC represents a pixel-wise comparison, while PRO focuses on region-based behavior. We show the best results for AUROC and PRO in bold. Remarkable, our approach is robust and represents state-of-the-art performance under both metrics.

![Figure 5. Histogram of anomaly scores for all categories of MVTec [3] (x-axis: anomaly score from 0 to 1, and y-axis: count).](image)

anomaly region.

Complexity analysis. Recent pre-trained model based approaches achieve promising performance by extracting features from anomaly-free samples as a measurement [7, 8]. However, storing feature models leads to large memory consumption. In comparison, our approach achieves better performance depending only on an extra CNN model. As shown in Tab. 3, Our model obtain performance gain with low time and memory complexity.

### Table 3. Comparison of pre-trained based approaches in terms of inference time (second on Intel i7), memory usage (MB), and performance (AD-AUROC/AL-AUROC/AL-PRO) on MVTec [3].

| Category/Method       | Methods       | Image Size | 128 | 256 |
|-----------------------|---------------|------------|-----|-----|
|                       |               |            |     |     |
| Textures              |               |            |     |     |
| Carpet                | SPADE (WResNet50) | 95.6/- | 98.195.3 |     |
| Grid                  | Ours          | 96.6/- | 96.6/96.6 | 95.0/- |
| Leather               | SPADE (WResNet50) | 98.1/- | 99.0/98.6 |     |
| Tile                  | Ours          | 98.4/- | 92.6/84.8 |     |
| Wood                  | SPADE (WResNet50) | 84.8/- | 92.1/82.3 |     |
|                       | Ours          | 96.3/- | 96.3/96.3 | 96.3/- |
| Objects               |               |            |     |     |
| Bottle                | SPADE (WResNet50) | 96.3/- | 98.294.7 |     |
| Cable                 | Ours          | 95.6/- | 98.2/97.0 | 96.0/- |
| Capsule               | SPADE (WResNet50) | 95.9/- | 96.9/87.2 |     |
| Hazelnut              | Ours          | 94.6/- | 98.8/89.2 |     |
| Metal Nut             | SPADE (WResNet50) | 86.4/- | 96.6/84.1 |     |
| Pill                  | Ours          | 86.4/- | 96.6/84.1 |     |
| Screw                 | SPADE (WResNet50) | 96.0/- | 98.3/94.4 |     |
| Toothbrush            | Ours          | 96.1/- | 98.2/86.7 |     |
| Transistor            | SPADE (WResNet50) | 76.5/- | 97.8/85.2 |     |
| Zipper                | Ours          | 93.9/- | 97.0/92.3 |     |
|                       |               |            |     |     |
| Average               |               |            |     |     |
| Total Average         |               |            |     |     |

Limitations. We observe that the localization performance on the transistor dataset is relatively weak, despite the good AD performance. This performance drop is caused by misinterpretation between prediction and annotation. As shown in Fig. 6, our method localize the misplaced regions, while the ground truth covers both misplaced and original areas. Alleviating this problem requires associating more features with contextual relationships. We empirically find that a higher-level feature layer with a wider perceptive field can improve the performance. For instance, anomaly detection with the second and third layer features achieves 94.5% AUROC, while using only the third layer improve the performance to 97.3%. In addition, reducing image resolution to 128 × 128 also achieves 97.6% AUROC. We present more cases of anomaly detection and localization, both positive and negative, in the supplementary material.

4.2. One-Class Novelty Detection

To evaluate the generality of proposed approach, we conduct one-class novelty detection experiment on 3 semantic datasets [32], MNIST [22], FashionMNIST [41] and CIFAR10 [20]. MNIST is a hand-written digits dataset from numbers 0-9. FashionMNIST consists of images from 10 fashion product classes. Both datasets includes 60K samples for training and 10K samples for testing, all in resolution of 28 × 28. CIFAR10 is a challenging dataset for novelty detection because of its inclusion of diverse natural objects. It includes 50K training images and 10K test images with scale of 32 × 32 in 10 categories.

Following the protocol mentioned in [27], we train the model with samples from a single class and detect novel samples. Note that the novelty score is defined as the sum.
Figure 6. Anomalies from top to bottom: "flip" on "metal nut", "misplaced" on "transistor" and "crack" on "hazelnut". Normal samples are provided as reference.

Figure 7. Visualization on tiny or inconspicuous anomalies. From left to right: carpet, tile, wood, cable, pill, toothbrush, and screw.

Table 4. AUROC(%) results for One-Class Novelty Detection. The best results are marked in bold.

| Method   | MNIST | F-MNIST | CIFAR10 | Caltech-256 |
|----------|-------|---------|---------|--------------|
| LSA [1]  | 97.5  | 92.2    | 64.1    | -            |
| OCGAN [27]| 97.3  | 87.8    | 65.7    | -            |
| HRN [17] | 97.6  | 92.8    | 71.3    | -            |
| DAAD [16]| 99.0  | -       | 75.3    | -            |
| MKD [33] | 98.7  | 94.5    | 84.5    | -            |
| G2D [28] | -     | -       | -       | 95.7         |
| OiG [45] | -     | -       | -       | 98.2         |
| Ours     | 99.3  | 95.0    | 86.5    | 99.9         |

Table 5. Ablation study on pre-trained bottleneck, OCE, and MFF.

| Backbone  | ResNet18 | ResNet50 | WResNet50 |
|-----------|----------|----------|-----------|
| AUROC AD  | 97.9     | 98.4     | 98.5      |
| AUROC AL  | 97.1     | 97.7     | 97.8      |
| RPO       | 91.2     | 93.1     | 93.9      |

Table 6. Quantitative comparison with different backbones.

5. Conclusion

We proposed a novel knowledge distillation paradigm, reverse distillation, for anomaly detection. It holistically addressed the problem in previous KD-based AD methods and boosted the T-S model’s response on anomalies. In addition, we introduced trainable one-class embedding and multi-scale feature fusion blocks in reverse distillation to improve one-class knowledge transfer. Experiments showed that our method significantly outperformed previous arts in anomaly detection, anomaly localization, and novelty detection.
References

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

[2] Samet Akcay, Amir Atapour-Abarghouei, and Toby P. Breckon. Anomaly: Semi-supervised anomaly detection via adversarial training. In C. V. Jawahar, Hongdong Li, Greg Mori, and Konrad Schindler, editors, Computer Vision – ACCV 2018, pages 622–637, Cham, 2019. Springer International Publishing.

[3] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Mvtec ad – a comprehensive real-world dataset for unsupervised anomaly detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2019.

[4] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2020.

[5] Paul Bergmann, Sindy Lüwe, Michael Fauser, David Sattlegger, and Carsten Steger. Improving unsupervised defect segmentation by applying structural similarity to autoencoders, 2018.

[6] Pengguang Chen, Shu Liu, Hengshuang Zhao, and Jiaya Jia. Distilling knowledge via knowledge review. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 5008–5017, 2021.

[7] Niv Cohen and Yedid Hoshen. Sub-image anomaly detection with deep pyramid correspondences, 2020.

[8] Thomas Defard, Aleksandr Setkov, Angelique Loesch, and Romaric Audigier. Padim: A patch distribution modeling framework for anomaly detection and localization. In International Conference on Pattern Recognition, pages 475–489. Springer, 2021.

[9] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In International Conference on Machine Learning, pages 1126–1135. PMLR, 2017.

[10] Izhak Golan and Ran El-Yaniv. Deep anomaly detection using geometric transformations. In Proceedings of the 32nd International Conference on Neural Information Processing Systems, NIPS’18, page 9781–9791, Red Hook, NY, USA, 2018. Curran Associates Inc.

[11] Dong Gong, Lingqiao Liu, Vuong Le, Budhadiya Saha, Moussa Reda Mansour, Svetla Venkatesh, and Anton van den Hengel. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), October 2019.

[12] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2, NIPS’14, page 2672–2680, Cambridge, MA, USA, 2014. MIT Press.

[13] Gregory Griffin, Alex Holub, and Pietro Perona. Caltech-256 object category dataset. 2007.

[14] 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 (CVPR), June 2016.

[15] Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. In NIPS Deep Learning and Representation Learning Workshop, 2015.

[16] Jialei Hou, Yingying Zhang, Qiaoyong Zhong, Di Xie, Shiliang Pu, and Hong Zhou. Divide-and-assemble: Learning block-wise memory for unsupervised anomaly detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pages 8791–8800, October 2021.

[17] Wenpeng Hu, Mengyu Wang, Qi Qin, Jinwen Ma, and Bing Liu. Hrn: A holistic approach to one class learning. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, Advances in Neural Information Processing Systems, volume 33, pages 19111–19124. Curran Associates, Inc., 2020.

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

[19] Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2013.

[20] Alex Krizhevsky. Learning multiple layers of features from tiny images, 2009.

[21] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1, NIPS’12, page 1097–1105, Red Hook, NY, USA, 2012. Curran Associates Inc.

[22] Yann LeCun. The mnist database of handwritten digits, 1998.

[23] Chun-Liang Li, Kihyuk Sohn, Jinsung Yoon, and Tomas Pfister. Cutpaste: Self-supervised learning for anomaly detection and localization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 9664–9674, June 2021.

[24] W. Liu, D. Lian W. Luo, and S. Gao. Future frame prediction for anomaly detection – a new baseline. In 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[25] Paolo Napoletano, Flavio Piccoli, and Raimondo Schettini. Anomaly detection in nanofibrous materials by cnn-based self-similarity. Sensors, 18(1), 2018.

[26] Hyunjong Park, Jongyoun Noh, and Bumsub Ham. Learning memory-guided normality for anomaly detection. In IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2020.

[27] Pramuditha Perera, Ramesh Nallapati, and Bing Xiang. Ocean: One-class novelty detection using gans with constrained
