

# 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 augmen-

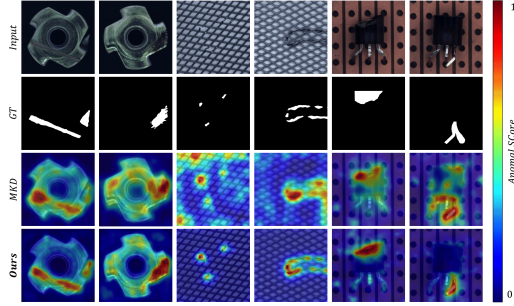


Figure 1. Anomaly detection examples on MVTec [3]. Multiresolution Knowledge Distillation (MKD) [33] adopts the conventional KD architecture in Fig. 2(a). Our reverse distillation method is capable of precisely localising a variate of anomalies.

tation [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