



Figure 2. T-S models and data flow in (a) conventional KD framework [6, 33] and (b) our *Reverse Distillation* paradigm.

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