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 discrimnating 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 $\mathcal{I}^t = \{I_1^t, ..., I_n^t\}$ be a set of available anomaly-free images and $\mathcal{I}^q = \{I_1^q, ..., I_m^q\}$ 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 $\mathcal{I}^t$ and $\mathcal{I}^q$ follow the same distribution. Out-of-distribution samples are considered anomalies.

System overview: Fig. 3 depicts the proposed reserve 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 $I \in \mathcal{I}^t$, the teacher E extracts multiscale 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 extreme 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 reserves 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 Ima-