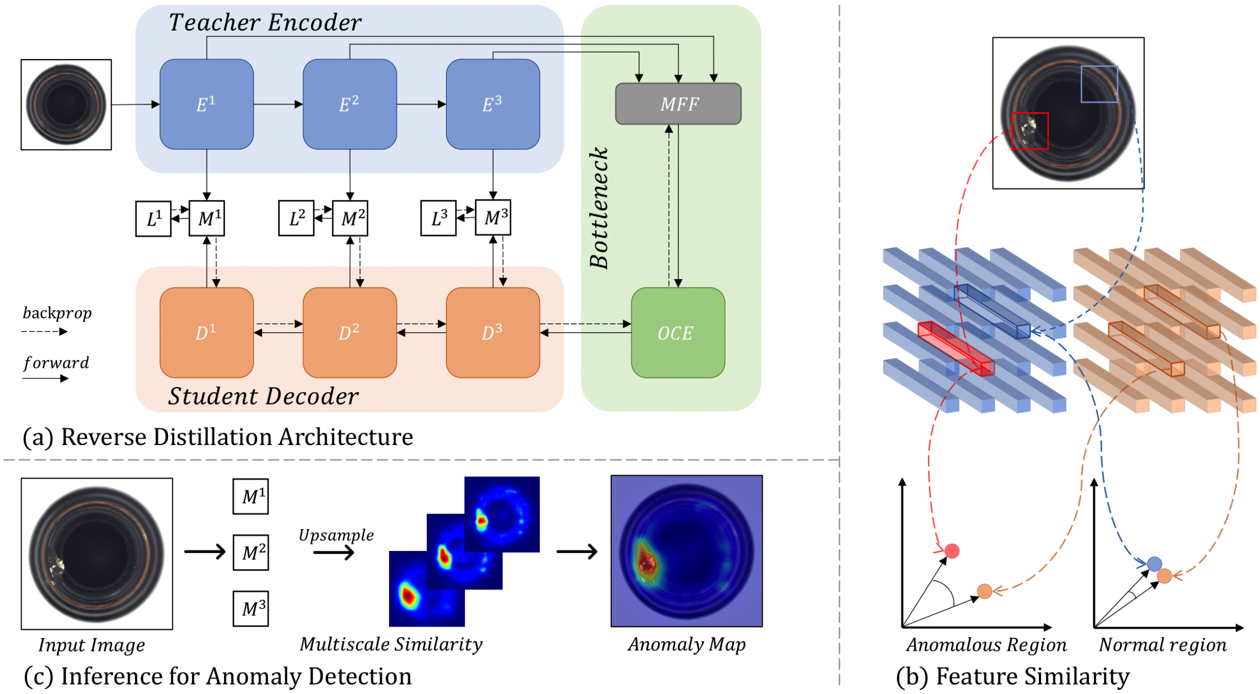

(a) Reverse Distillation Architecture

(c) Inference for Anomaly Detection

(b) Feature Similarity

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 $\mathcal{L}$. (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.

geNet [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 ϕ 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 $M^k \in \mathbb{R}^{H_k \times W_k}$:

$$M^k(h,w) = 1 - \frac{(f_E^k(h,w))^T \cdot f_D^k(h,w)}{\left\|f_E^k(h,w)\right\| \left\|f_D^k(h,w)\right\|}. \qquad (1)$$