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:

$$\mathcal{L}_{\mathcal{KD}} = \sum_{k=1}^{K} \left\{ \frac{1}{H_k W_k} \sum_{h=1}^{H_k} \sum_{w=1}^{W_k} M^k(h,w) \right\}, \qquad (2)$$

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 decoder 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 residule 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 downsample the shallow features through one or more 3×3 convolutional layers with stride of 2, followed by batch normalization and ReLU activation function. Then a 1×1 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.

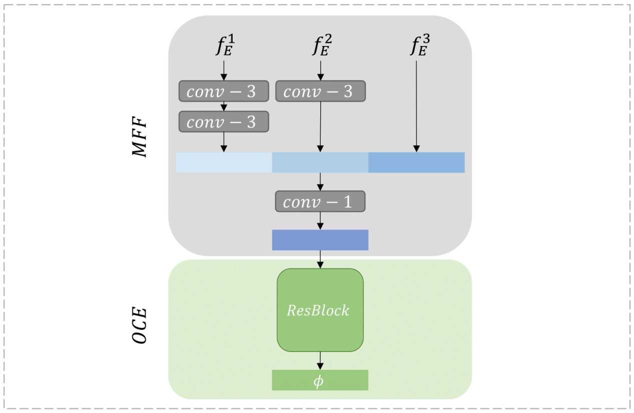

Figure 4. Our one-class bottleneck embedding module consists of trainable MFF and OCE blocks. MFF aligns multi-scale features from teacher E and OCE condenses the obtained rich feature to a compact bottleneck code ϕ.

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 Ψ denotes the bilinear up-sampling operation used in this study. Then a precise score map S_{I^q} is formulated as the pixel-wise accumulation of all anomaly maps:

$$S_{AL} = \sum_{i=1}^{L} \Psi(M^i). \qquad (3)$$

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