

Reconstruction-based approaches assume that the reconstruction model trained with normal samples has a generalization gap with anomalies, thus fails to reconstruct anomalies. AE [6,13,16,25] and GAN [26,29,39] are intuitive choices of reconstruction models. Zhou et al. [40] and Xia et al. [35] respectively adopt the structural information and semantic segmentation information for better reconstruction. Zaheer et al. [39] utilize a discriminator to distinguish good or bad quality of reconstruction, and the predicted possibility of bad quality serves as an anomaly score. Gong et al. [16] and Park et al. [25] introduce a memory module to select the most similar embedding in embedding storage of normal samples to restrict the generalization on anomalies. Dehaene et al. [12] refine the selection method with an iterative gradient-based approach.

Projection-based approaches project samples into an embedding space, where normal samples and anomalies are more distinguishable. SVDD [28] extracts feature representation with the one-class classification objective. Yi and Yoon [37] propose a patch-based SVDD with multiple kernels. Liu et al. [21] and Kwon et al. [19] find that the back-propagated gradients of normal samples and anomalies are more distinguishable. FCDD [22] is trained to enlarge the embedding differences between normal samples and anomalies, where the mapped samples are themselves an explanation heat map. Bergmann et al. [5] utilize a teacher-student network, assuming that the embedding differences between normal samples and anomalies would be enlarged through knowledge distillation. Salehi et al. [30] extend the knowledge distillation to multi-layer, multi-scale scheme, enlarging the distillation gap between normal samples and anomalies. PaDiM [11] models normal distribution using pre-trained features, then utilize a distance metric to measure the anomalies. Wang et al. [34] compare the embeddings of local pattern and global pattern to detect anomalies.

Transformer in anomaly detection. Transformer [33] has been successfully used in computer vision [9]. Some attempts also try to utilize transformer for anomaly detection. InTra [27] adopts transformer to recover the image by recovering all masked patches one by one. VT-ADL [24] and AnoVit [38] both apply transformer encoder to reconstruct images. However, these methods mainly focus on indistinguishable raw pixels, and do not figure out why transformer brings improvement. In contrast, we reconstruct pre-trained features instead of raw pixels. We also confirm the efficacy of the query embedding in attention layer to prevent the “identical shortcut”.

3 Method

In this part, we first introduce the architecture of ADTR, followed by the analysis of why transformer could limit to reconstruct anomalies well. Finally, we propose two loss functions to extend our approach compatible with available anomalies.

3.1 Architecture

Embedding. A frozen pre-trained CNN backbone is first utilized for feature extraction (Fig. 2a). Here we use EfficientNet-B4 [32] pre-trained on ImageNet.