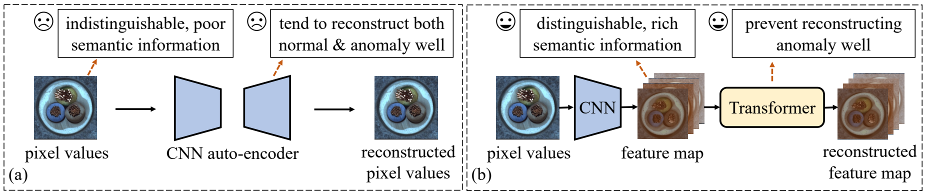
☹ indistinguishable, poor semantic information
☹ tend to reconstruct both normal & anomaly well
pixel values
CNN auto-encoder
reconstructed pixel values
(a)
☺ distinguishable, rich semantic information
☺ prevent reconstructing anomaly well
pixel values
CNN
feature map
Transformer
reconstructed feature map
(b)

Fig. 1: **(a) CNN-based pixel reconstruction methods** tend to reconstruct both normal samples and anomalies well, making them still hard to distinguish. Also, the pixel values contain indistinguishable semantic information. **(b) Our method** reconstructs features with distinguishable semantic information. Besides, the adoption of transformer limits the reconstruction of anomalies.

As show in Fig. 1a, one concern about these approaches is the poor representation ability. The reconstruction targets are raw pixel values with poor semantic information. Therefore, these pixel reconstruction approaches usually fails when normal and anomalous regions share similar pixel values but different semantic information like different textures. In another aspect, it has been verified that the feature extractor pre-trained on large public datasets could extract distinguishable features for normal samples and anomalies [5,30]. Thus we propose to reconstruct pre-trained features instead of raw pixel values.

Taking CNN as the reconstruction model brings another issue (Fig. 1a). CNN tends to take shortcuts to learn a somewhat "identical mapping", which means the anomalous regions are also reconstructed quite well [16]. The great success of transformer in computer vision inspires us to propose a transformer-based reconstruction model. The query embedding in attention layer of transformer could limit the tendency of "identical mapping", which helps distinguish normal samples and anomalies (See Sec. 3.2).

Besides, more anomaly samples are available with the runs of production lines [5], bringing anomaly detection the demands of compatibility with both the normal-sample-only case (only normal samples are available) and the anomaly-available case (normal samples and a few anomalies are available). Therefore, a unified approach that is compatible with both cases would be a better solution.

In this paper, we propose a concise but powerful transformer-based anomaly detection approach. As shown in Fig. 1b, a frozen pre-trained CNN backbone is adopted to extract features, then a transformer is used for feature reconstruction. The proposed approach has strong representation abilities, and could limit the tendency of "identical mapping". Moreover, novel loss functions are proposed for the compatibility with the anomaly-available case. The performance could be further improved by adding simple synthetic or external irrelevant anomalies. Our approach achieves state-of-the-art anomaly detection performance in anomaly detection datasets including MVTec-AD [4] and CIFAR-10 [18].

2 Related Work

Existing anomaly detection approaches could be generally divided into two categories: reconstruction-based ones and projection-based ones.