



Fig. 4: **Anomaly detection results on MVTec-AD** [4]. From left to right: the anomaly sample, the ground-truth, and the anomaly score map of ADTR.

samples for training, and test on both normal and anomaly samples. In *anomaly-available case*, following [22], we synthesize anomalies by adding confetti noise on normal samples (Fig. 3).

CIFAR-10 [18] is a classical classification dataset with 10 classes. Each class has 5000 images for training and 1000 images for testing. In *normal-sample-only case*, following [19], the training set of one class is used for training, and the test set contains normal images of the same class and the same number of anomaly images randomly sampled from other classes. In *anomaly-available case*, an irrelevant dataset, CIFAR-100 [18], is used as an auxiliary dataset. We randomly select the same number of images from CIFAR-100 as anomalies.

4.2 Anomaly Detection on MVTec-AD

The performance of our method is evaluated on anomaly detection and localization tasks of MVTec-AD [4].

Setup. The sizes of the image and feature map are selected as 256×256 and 16×16 , respectively. The numbers of the encoder layer and decoder layer (N in Fig. 2) in transformer are both set as 4. The features from *layer1* to *layer5* of EfficientNet-B4 [32] are resized and concatenated to form a 720-channel feature map. The reduced channel dimension is set as 256. AdamW optimizer [23] with weight decay 1×10^{-4} is used for training with batch size 16. In *normal-sample-only case*, models are trained with \mathcal{L}_{norm} in Eq. (1) for 500 epochs. The learning rate is 1×10^{-4} initially, and dropped by 0.1 after 400 epochs. In *anomaly-available case*, the pixel-level loss, \mathcal{L}_{px} , in Eq. (7) is adopted for training, where α is chosen as 0.003. The trained model in normal-sample-only case is firstly loaded. Then the model is trained for 300 epochs with the learning rate of 1×10^{-4} for first 200 epochs and 1×10^{-5} for last 100 epochs.