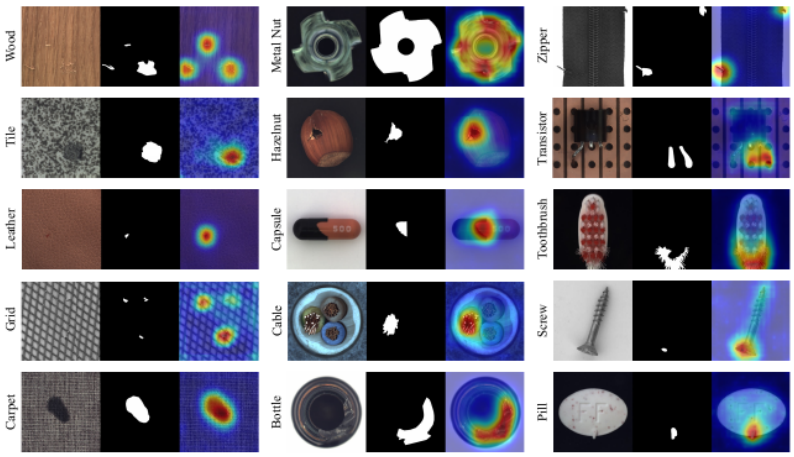

図4：MVTec-AD [4] における異常検出結果。左から右へ：異常サンプル、真値、およびADTRの異常スコアマップ。

トレーニングには正常サンプルのみを使用し、テストには正常サンプルと異常サンプルの両方を使用します。異常サンプルが利用可能な場合、[22]に従い、正常サンプルにコンフェッティノイズを追加して異常を合成します（図3）。
CIFAR-10 [18]は10クラスの古典的な分類データセットです。各クラスにはトレーニング用に5000枚、テスト用に1000枚の画像が含まれます。正常サンプルのみの場合、[19]に従い、1クラスのトレーニングセットをトレーニングに用い、テストセットには同じクラスの正常画像と、他のクラスからランダムにサンプリングされた同じ数の異常画像を含めます。異常データが利用可能な場合、関連しないデータセットであるCIFAR-100 [18]を補助データセットとして使用します。CIFAR-100から異常画像として同じ数の画像をランダムに選択します。

4.2 MVTec-AD における異常検出

当該手法の性能は、MVTec-AD [4]の異常検出と局所化タスクで評価されます。

設定。画像と特徴マップのサイズはそれぞれ256 × 256と16 × 16に設定されます。トランスフォーマーのエンコーダー層とデコーダー層（図2のN）の層数はどちらも4に設定されます。EfficientNet-B4 [32]の層1から層5の特徴はリサイズされ、結合されて720チャネルの特徴マップを形成します。チャネル次元は256に設定されます。トレーニングにはバッチサイズ16で、重み減衰1 × 10^{-4}のAdamWオプティマイザー[23]が使用されます。正常サンプルのみの場合、式(1)のL_{norm}で500エポックトレーニングされます。学習率は初期値として1 × 10^{-4}に設定され、400エポック後に0.1ずつ減少されます。異常データ利用可能な場合、式(7)のピクセルレベル損失L_{px}がトレーニングに採用され、αは0.003に設定されます。正常サンプルのみの場合に訓練されたモデルを最初に読み込みます。その後、最初の200エポックで学習率1×10^{-4}、最後の100エポックで1 × 10^{-5}を使用して、300エポック間訓練します。

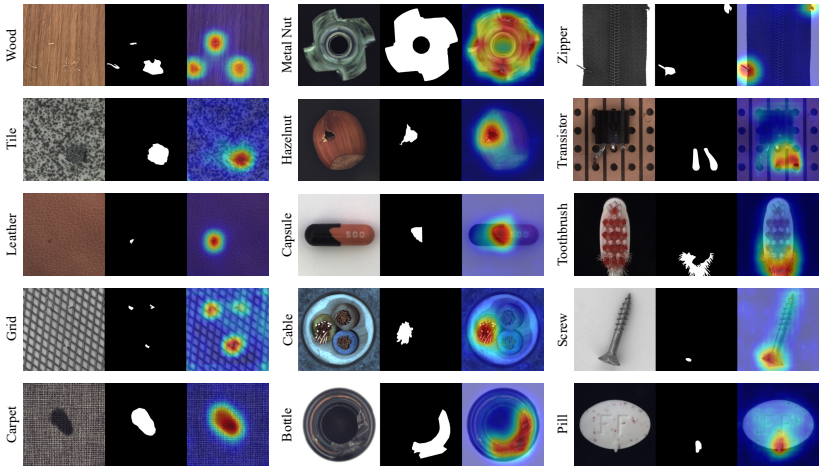

Fig. 4: **Anomaly detection results on MVTec-AD** [4]. From left to right: the anomaly sample, the ground-truth, and the anomaly score map of ADTR.

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.