再構築ベースのアプローチは、正常なサンプルで訓練された再構築モデルが異常値に対して汎化ギャップを有し、異常値の再構築に失敗すると仮定しています。AE [6,13,16,25]とGAN [26,29,39]は、再構築モデルの直感的な選択です。Zhouら [40]とXiaら [35]はそれぞれ、構造情報とセマンティックセグメンテーション情報を用いて再構築の精度を向上させています。Zaheerら[39]は、再構築の品質を良いか悪いか区別するディｃｋｐиミネーターを利用し、予測された不良品質の可能性が異常スコアとして機能します。Gongら[16]とParkら[25]は、異常に対する汎化を制限するため、正常なサンプルの埋め込みストレージから最も類似した埋め込みを選択するメモリモジュールを導入しています。Dehaeneら[12]は、反復的な勾配ベースのアプローチで選択方法を精緻化しています。
投影ベースのアプローチは、サンプルを埋め込み空間に投影し、正常なサンプルと異常なサンプルをより明確に区別できるようにします。SVDD [28] は、1クラス分類の目的関数を用いて特徴表現を抽出します。Yi と Yoon [37] は、複数のカーネルを用いたパッチベースのSVDDを提案しています。Liuら[21]とKwonら[19]は、正常なサンプルと異常なサンプルのバックプロパゲーション勾配がより区別しやすくなることを発見しました。FCDD [22]は、正常なサンプルと異常なサンプルの埋め込み差を拡大するように訓練され、マッピングされたサンプル自体が説明用ヒートマップとなります。Bergmannら[5]は、教師-生徒ネットワークを活用し、知識蒸留を通じて正常サンプルと異常サンプルの埋め込み差を拡大すると仮定しています。Salehiら[30]は知識蒸留を多層・多スケール方式に拡張し、正常サンプルと異常サンプルの蒸留ギャップを拡大しています。PaDiM[11]は事前学習済み特徴量で正常分布をモデル化し、距離メトリクスの測定で異常を検出します。Wangら[34]は、局所パターンとグローバルパターンの埋め込みを比較して異常を検出しています。

異常検出におけるトランスフォーマー。トランスフォーマー[33]はコンピュータビジョン[9]で成功裡に活用されています。一部の試みでは、トランスフォーマーを異常検出に活用する試みも行われています。InTra[27]は、マスクされたパッチを一つずつ復元することで画像を復元するためにトランスフォーマーを採用しています。VT-ADL [24]とAnoVit [38]は、トランスフォーマーエンコーダーを用いて画像の再構築を行っています。しかし、これらの方法は主に区別できない raw ピクセルに焦点を当てており、トランスフォーマーが改善をもたらす理由を明確に説明していません。これに対し、私たちは raw ピクセルではなく事前学習済み特徴量を再構築します。また、アテンション層におけるクエリ埋め込みの有効性を確認し、「同一ショートカット」を防止します。

3 Method

このセクションでは、まずADTRのアーキテクチャを紹介し、次にトランスフォーマーが異常の再構築に限定される理由を分析します。最後に、既存の異常に対応可能なアプローチを拡張するための2つの損失関数を提案します。

3.1 Architecture

埋め込み。最初に凍結された事前訓練済みCNNバックボーンを特徴抽出に利用します（図2a）。ここでは、ImageNetで事前訓練されたEfficientNet-B4[32]を使用しています。

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.