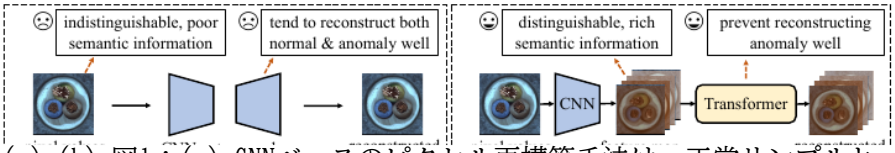

(a) (b) 図1：(a) CNNベースのピクセル再構築手法は、正常サンプルと異常を同様に再構築するため、依然として区別が困難です。また、ピクセル値には区別不能なセマンティック情報が含まれます。(b) 当社の手法は、区別可能なセマンティック情報を含む特徴を再構築します。さらに、トランスフォーマーの採用により、異常の再構築が制限されます。

図1aに示すように、これらのアプローチの課題の一つは、表現能力の低さです。再構築の対象は、意味的な情報が乏しい raw ピクセル値です。したがって、正常領域と異常領域が類似したピクセル値を持つが、異なる意味的な情報（例えば異なるテクスチャ）を持つ場合、これらのピクセル再構築アプローチは通常失敗します。別の観点から、大規模な公開データセットで事前学習された特徴抽出器は、正常サンプルと異常サンプルに対して区別可能な特徴を抽出できることが確認されています [5,30]。したがって、私たちは生ピクセル値ではなく事前学習済み特徴を再構築することを提案します。

CNNを再構築モデルとして採用すると、別の問題が生じます（図1a）。CNNは「同一マッピング」を学習する傾向があり、異常領域も比較的よく再構築されます[16]。コンピュータビジョンにおけるトランスフォーマーの大きな成功は、私たちにトランスフォーマーベースの再構築モデルを提案するきっかけとなりました。トランスフォーマーの注意層におけるクエリ埋め込みは、「同一のマッピング」傾向を抑制し、正常サンプルと異常を区別するのに役立ちます（セクション3.2参照）。

さらに、生産ラインの稼働に伴い、より多くの異常サンプルが利用可能になります[5]。これにより、異常検出は「正常サンプルのみの場合」（正常サンプルのみが利用可能）と「異常サンプルが利用可能な場合」（正常サンプルと少数の異常サンプルが利用可能）の両方に対応する柔軟性が求められます。したがって、両ケースに対応可能な統一的なアプローチがより適切な解決策となります。
本論文では、簡潔ながら強力なトランスフォーマーベースの異常検出アプローチを提案します。図1bに示すように、凍結された事前訓練済みCNNバックボーンを採用して特徴量を抽出後、トランスフォーマーを用いて特徴再構築を行います。提案手法は強力な表現能力を有し、「同一マッピング」の傾向を抑制できます。さらに、異常データが存在するケースとの互換性を考慮した新たな損失関数を提案しています。単純な合成異常データや外部からの関連性のない異常データを追加することで、性能をさらに向上させることができます。当アプローチは、MVTec-AD [4]やCIFAR-10 [18]を含む異常検出データセットにおいて、最先端の異常検出性能を達成しています。

2 関連研究

既存の異常検出アプローチは、一般的に再構築ベースと投影ベースの2つのカテゴリーに分類できます。

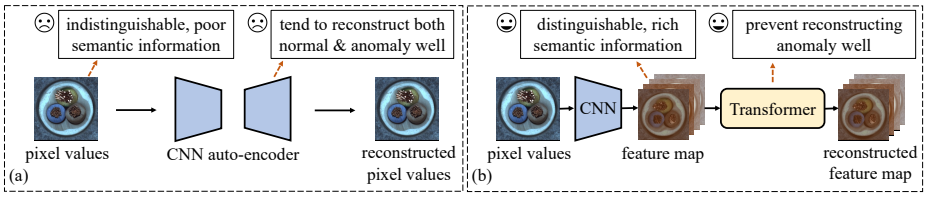

Fig. 1: **(a) CNN-based pixel reconstruction methods** tend to reconstruct both normal samples and anomalies well, making them still hard to distinguish. Also, the pixel values contain indistinguishable semantic information. **(b) Our method** reconstructs features with distinguishable semantic information.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