MVTec-ADにおけるトレーニング構成。正常サンプルのみの場合、バックボーンは固定されます。トランスフォーマーは式(3)のL_{norm}を用いてバッチサイズ16で500エポックトレーニングされます。AdamWオプティマイザー[23]（ウェイトデカay 1 × 10^{-4}）が使用されます。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。異常データが利用可能な場合、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。トランスフォーマーは式(6)のL_{px}を用いて300エポック訓練されます。式(6)のαは0.003に設定されます。学習率は初期値として1 × 10^{-4}に設定され、200エポック後に0.1ずつ減少させます。
CIFAR-10におけるトレーニング構成。正常サンプルのみの場合、バックボーンで説明された画像サイズと特徴量サイズを除き、MVTec-ADの場合と同一の設定です。より効率的なトレーニングのため、バッチサイズは128に設定されます。異常データが利用可能な場合、MVTec-ADと同じ実装を採用しますが、以下の点のみ異なります。CIFAR-10の場合、異常は画像レベルでラベル付けされているため、トランスフォーマーは式(8)のL_{img}で訓練されます。ここで、αとkはそれぞれ0.003と20に設定されます。

B 追加の可視化結果

MVTec-ADにおける定性的な結果が示されています。これらのカテゴリーには、以下のものが含まれます：カーペット（図A1）、グリッド（図A2）、レザー（図A3）、タイル（図A4）、木材（図A5）、ボトル（図A6）、ケーブル（図A7）、カプセル（図A8）、ヘーゼルナッツ（図A9）、 金属ナット（図A10）、錠剤（図A11）、ネジ（図A12）、歯ブラシ（図A13）、トランジスタ（図A14）、ジッパー（図A15）。当アプローチは、すべてのカテゴリにおいて多様な異常を高い局所化精度で検出可能です。提案されたアプローチの性能は、これらのカテゴリにおける多様な異常タイプにおいて安定しており、強い汎化能力と頑健性を示しています。具体的には、非常に小さな異常（例：図A8の2列目）と非常に大きな異常（例：図A4の9列目）、単一種類の異常（例：図A3の2列目）と複数種類の組み合わせ異常（例：図A5の最終列）、テクスチャまたは色の乱れ（例：図A1の2列目）と配置の誤り（例：図A14の最終列）において、当アプローチはすべての異常を効果的に検出できました。

CIFAR-10における定性的結果を示します。これらのカテゴリには、飛行機（図A16）、自動車（図A17）、鳥（図A18）、猫（図A19）、鹿（図A20）、犬（図A21）、カエル（図A22）、馬（図A23）、船（図A24）、トラック（図A25）が含まれます。当社のアプローチは、さまざまな種類の異常を成功裡に検出できました。また、異常スコアの高さは主に異常対象物に集中しており、背景ではなく、これは当社のアプローチがセマンティック特徴の理解に基づいて異常を検出していることを示しています。特に、正常サンプルと非常に類似した異常の場合でも、例えば「自動車」カテゴリが正常サンプルとして機能する「トラック」カテゴリ（図A17の6列目）や、「鹿」カテゴリが正常サンプルとして機能する「犬」 カテゴリーが正常サンプルとして「猫」カテゴリーが使用される場合（例えば図A19の最終列）、または「馬」カテゴリーが正常サンプルとして「鹿」カテゴリーが使用される場合（例えば図A20の10列目）でも、当社のアプローチはこれらの異常を正常サンプルから成功裏に区別しています。

Training configurations on MVTec-AD. In *normal-sample-only case*, the backbone is frozen. The transformer is trained with $\mathcal{L}_{norm}$ in Eq. (3) for 500 epochs with batch size 16. AdamW optimizer [23] with weight decay 1×10^{-4} is used. The learning rate is set as 1×10^{-4} initially, and dropped by 0.1 after 400 epochs. In *anomaly-available case*, the trained model in *normal-sample-only case* is firstly loaded. The transformer is trained with $\mathcal{L}_{px}$ in Eq. (6) for 300 epochs. α in Eq. (6) is set as 0.003. The learning rate is initially set as 1×10^{-4}, and dropped by 0.1 after 200 epochs.

Training configurations on CIFAR-10. In *normal-sample-only case*, the details are the same as those in **MVTec-AD** except the image size and feature size described in **Backbone**. For more efficient training, the batch size is set as 128. In *anomaly-available case*, the same implementations as **MVTec-AD** are adopted except the followings. Considering that the anomalies are image-level labeled in CIFAR-10 case, the transformer is trained with $\mathcal{L}_{img}$ in Eq. (8), where α and k are selected as 0.003 and 20, respectively.

B More Visualization Results

Qualitative results on MVTec-AD are provided. These categories include: carpet (Fig. A1), grid (Fig. A2), leather (Fig. A3), tile (Fig. A4), wood (Fig. A5), bottle (Fig. A6), cable (Fig. A7), capsule (Fig. A8), hazelnut (Fig. A9), metal nut (Fig. A10), pill (Fig. A11), screw (Fig. A12), toothbrush (Fig. A13), transistor (Fig. A14), and zipper (Fig. A15). Our approach could detect different kinds of anomalies in all categories with quite high localization accuracy. The performance of the proposed approach keeps stable in all these categories with various anomaly types, demonstrating strong generalization ability and robustness. Specifically, for both quite small anomalies (e.g. the second column in Fig. A8) and quite large anomalies (e.g. the ninth column in Fig. A4), both single-kind anomalies (e.g. the second column in Fig. A3) and multi-kind combined anomalies (e.g. the last column in Fig. A5), both texture or color disorder (e.g. the second column in Fig. A1) and misplacement (e.g. the last column in Fig. A14), our approach could effectively detect all anomalies.

Qualitative results on CIFAR-10 are given. These categories include: airplane (Fig. A16), automobile (Fig. A17), bird (Fig. A18), cat (Fig. A19), deer (Fig. A20), dog (Fig. A21), frog (Fig. A22), horse (Fig. A23), ship (Fig. A24), and truck (Fig. A25). Our approach could successfully detect various kinds of anomalies. Also, high anomaly scores mainly center on the anomaly objects rather than the backgrounds, which indicates that our approach detects anomalies based on the understanding of semantic features. In particular, even for anomalies that are very similar to normal samples, like the "truck" category when "automobile" category serves as normal samples (e.g. the sixth column in Fig. A17), the "dog" category when "cat" category serves as normal samples (e.g. the last column in Fig. A19), the "horse" category when "deer" category serves as normal samples (e.g. the tenth column in Fig. A20), our approach still successfully distinguishes these anomalies from normal samples.