

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.