Table 4: **Ablation study** on (a) attention & auxiliary query embedding, (b) reconstructing pixels *vs.* features, (c) backbone, and (d) multi-scale features under pixel-level AUROC metric on anomaly localization of MVTec-AD [4].

(a) Attention & auxiliary query embedding

	CNN	w/o Attn	w/o Query	Attn+Query
Pixel AUROC	94.4	94.8	94.2	**97.2**

(b) Reconstructing pixels *vs.* features

	Pixels	Features
Pixel AUROC	91.3	**97.2**

(c) Backbone

	Res-18	Res-34	Efficient-B0	Efficient-B4
Pixel AUROC	95.3	95.7	96.4	**97.2**

(d) Multi-scale features

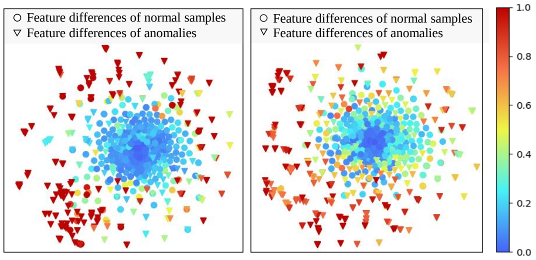
	Last-layer	Multi-scale
Pixel AUROC	96.0	**97.2**

Reconstructed target. In Tab. 4b, reconstructing features surpasses pixel values substantially, indicating that the features extracted by pre-trained backbone are more distinguishable for normal samples and anomalies than raw pixels.

Backbone and multi-scale features. (1) As shown in Tab. 4c, four different backbones all achieve quite good performance, reflecting that our method could cooperate with different types of backbones. (2) In Tab. 4d, multi-scale features obviously outperform last-layer feature, because multi-scale features contain different levels of receptive fields thus are sensitive to different anomalies.

4.5 Visualization of Feature Difference Vectors

We visualize the feature difference vectors $\boldsymbol{d}(:, u)$ in Eq. (2) to better interpret our approach. Specifically, we randomly sample 600 feature difference vectors (normal : anomaly = 1:1) from MVTec-AD [4]. Then t-SNE is utilized to visualize the high dimensional vectors in a 2D space, as shown in Fig. 5. Firstly, normal samples and anomalies are mostly colored with blue and red, respectively, indicating good anomaly detection ability. Secondly, normal samples are well clustered, and there is a wide gap between the normal samples and anomalies. These observations indicate that our approach brings a large generalization gap between normal samples and anomalies.



(a) Texture Feature Differences (b) Object Feature Differences

Fig. 5: **visualization of feature difference vectors** by t-SNE. Circles and triangles respectively represent normal samples and anomalies. The color map indicates the predicted anomaly possibility. Our method brings large generalization gap between normal samples and anomalies.

5 Conclusion

In this paper, we propose anomaly detection transformer to utilize a transformer to reconstruct pre-trained features. First, the pre-trained features contain dis-