Table 2: **Anomaly detection results under image-level AUROC metric on MVTec-AD [4].**

	Texture					Object										
	Carp.	Grid	Leat.	Tile	Wood	Bott.	Cable	Caps.	Haze.	Meta.	Pill	Screw	Toot.	Tran.	Zipp.	Mean
GANomaly [2]	69.9	70.8	84.2	79.4	83.4	89.2	75.7	73.2	78.5	70.0	74.3	74.6	65.3	79.2	74.5	76.2
SCADN [36]	50.4	98.3	65.9	79.2	96.8	95.7	85.6	76.5	83.3	62.4	81.4	83.1	98.1	86.3	84.6	81.8
ARNet [14]	70.6	88.3	86.2	73.5	92.3	94.1	83.2	68.1	85.5	66.7	78.6	**100**	**100**	84.3	87.6	83.9
SPADE [10]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	85.5
KDAD [30]	79.3	78.0	95.1	91.6	94.3	99.4	89.2	80.5	98.4	73.6	82.7	83.3	92.2	85.6	93.2	87.7
PSVDD [37]	98.6	90.3	76.7	92.9	94.6	92.0	90.9	**94.0**	86.1	81.3	**97.8**	**100**	91.5	96.5	**97.9**	92.1
TS [5]	95.3	**98.7**	93.4	95.8	95.5	96.7	82.3	92.8	91.4	94.0	86.7	87.4	98.6	83.6	95.8	92.5
ADTR(ours)	**100**	97.5	**100**	**100**	**99.9**	**100**	**92.5**	93.1	**100**	**94.9**	92.1	94.0	93.1	97.6	95.8	96.4
ADTR+(ours)	**100**	97.8	**100**	**100**	**99.9**	**100**	**92.5**	92.5	99.9	94.5	93.3	94.2	93.9	**98.0**	97.0	**96.9**

Table 3: **Anomaly detection results under image-level AUROC metric on CIFAR-10 [18].**

	Airplane	Automobile	Bird	Cat	Deer	Dog	Frog	Horse	Ship	Truck	Mean
VAE [3]	63.4	44.2	64.0	49.7	74.3	51.5	74.5	52.7	67.4	41.6	58.3
KDE [7]	65.8	52.0	65.7	49.7	72.7	49.6	75.8	56.4	68.0	54.0	61.0
AnoGAN [31]	67.1	54.7	52.9	54.5	65.1	60.3	58.5	62.5	75.8	66.5	61.8
LSA [1]	73.5	58.0	69.0	54.2	76.1	54.6	75.1	53.5	71.7	54.8	64.1
DSVDD [28]	61.7	65.9	50.8	59.1	60.9	65.7	67.7	67.3	75.9	73.1	64.8
OCGAN [26]	75.7	53.1	64.0	62.0	72.3	62.0	72.3	57.5	82.0	55.4	65.7
GradCon [19]	76.0	59.8	64.8	58.6	73.3	60.3	68.4	56.7	78.4	67.8	66.4
Loc-Glo [34]	79.1	70.3	67.5	56.1	73.9	63.8	73.2	67.4	81.4	72.2	70.5
TS [5]	78.9	84.9	73.4	74.8	85.1	79.3	89.2	83.0	86.2	84.8	82.0
GT [15]	76.2	84.8	77.1	73.2	82.8	84.8	82.0	88.7	89.5	83.4	82.3
KDAD [30]	90.5	90.4	80.0	77.0	86.7	91.4	89.0	86.8	91.5	88.9	87.2
ADTR(ours)	94.1	97.4	92.3	89.0	93.2	94.4	97.4	95.8	96.3	96.7	94.7
ADTR+(ours)	**96.2**	**98.0**	**94.5**	**91.7**	**95.1**	**95.6**	**98.0**	**97.1**	**98.0**	**96.9**	**96.1**

anomaly-available case, the performance of ADTR+ is further improved by 1.4% with the help of external irrelevant dataset, reflecting the effectiveness of the designed image-level loss function, $\mathcal{L}_{img}$.

4.4 Ablation Study

Extensive ablation studies with pixel-level AUROC metric are conducted on anomaly localization task of MVTec-AD [4].

Attention and auxiliary query embedding. As shown in Tab. 4a, a CNN revised from ResNet [17] is firstly included as the baseline of the reconstruction model. (1) The replacement of the attention layer is a concatenation followed by projection. If we remove the attention layer (w/o Attn) from the transformer, the performance shows no obvious superiority to CNN. (2) Without the auxiliary query embedding (w/o Query), meaning that only the encoder embedding is input to the decoder, the performance is even worse than CNN. (3) Equipped with both attention and auxiliary query embedding (Attn+Query), transformer stably outperforms CNN by 2.8%. This proves our assertion in Sec. 3.2 that the auxiliary query embedding in attention layer helps prevent transformer from reconstructing anomalies well.