

Table 7: The average performance of all methods is evaluated under two backbone networks. "S2S" denotes the Synthetic Benchmark, "S2R" represents the Synth to Real Benchmark, and "R2R" indicates the Real to Real Benchmark.

Method	AUROC↑				FPR95↓			
	S2S	S2R	R2R	Avg	S2S	S2R	R2R	Avg
MSP[14]	83.3	71.2	75.8	76.8	67.1	86.9	80.0	78.0
MLS[56]	83.9	70.3	77.8	77.3	56.8	86.1	72.4	71.8
ODIN[27]	83.4	70.9	78.0	77.4	57.8	86.7	71.4	72.0
Energy[29]	84.1	70.2	77.6	77.3	56.8	86.6	72.7	72.0
GradNorm[17]	67.8	68.2	73.8	69.9	80.0	87.4	76.1	81.2
ReAct[50]	87.3	70.1	78.9	78.8	48.4	86.0	71.4	68.6
OE+mixup[15]	82.7	65.5	69.4	72.5	63.2	91.2	85.8	80.1
Cosine Proto	86.7	68.1	80.8	78.5	52.3	86.6	68.8	69.2
CE(L^2)	88.9	71.9	77.0	79.3	43.9	85.8	75.1	68.3
SubArcFace[10]	86.5	74.3	78.3	79.7	53.1	85.2	73.1	70.5
NF	84.9	73.9	75.8	78.2	56.7	83.3	74.0	71.3
VAE[35]	77.0	63.3	54.8	65.0	64.2	84.6	92.8	80.5
VAE(Ours)	87.9	67.4	80.6	78.3	50.5	82.9	79.8	71.1

A AVERAGE PERFORMANCE ACROSS TWO BACKBONE

To further investigate the stability of the methods evaluated in 3DOS, we computed their average performance across two backbone networks and presented them in Tab. 7. We specifically included methods that underwent experimentation across all three benchmarks. It can be observed that our method performs well on the *Synthetic Benchmark*, with AUROC and FPR95 metrics closely approaching the SOTA and ranking second in stability. However, its performance on the subsequent two benchmarks is less satisfactory, likely due to the challenges posed by real point cloud data. We plan to further investigate this issue in future research.