Contributions of BSD at 1.0 Detection Threshold CWAttack on MNIST 100% 80% Classification accuracy 60% 40% MagNet Accuracy 20% **BSD** Accuracy No Defense Accuracy 0% 30 10 20

Confidence in CWAttack