Confidence EAD Attack against BSD at 10.0% Detection Threshold CIFAR 100% 80% Classification accuracy 60% 40% MagNet BSD with JSD 20% BSD with Confidence BSD w/o Confidence and JSD 0% 20 60 40 80 100

Confidence in EAD attack