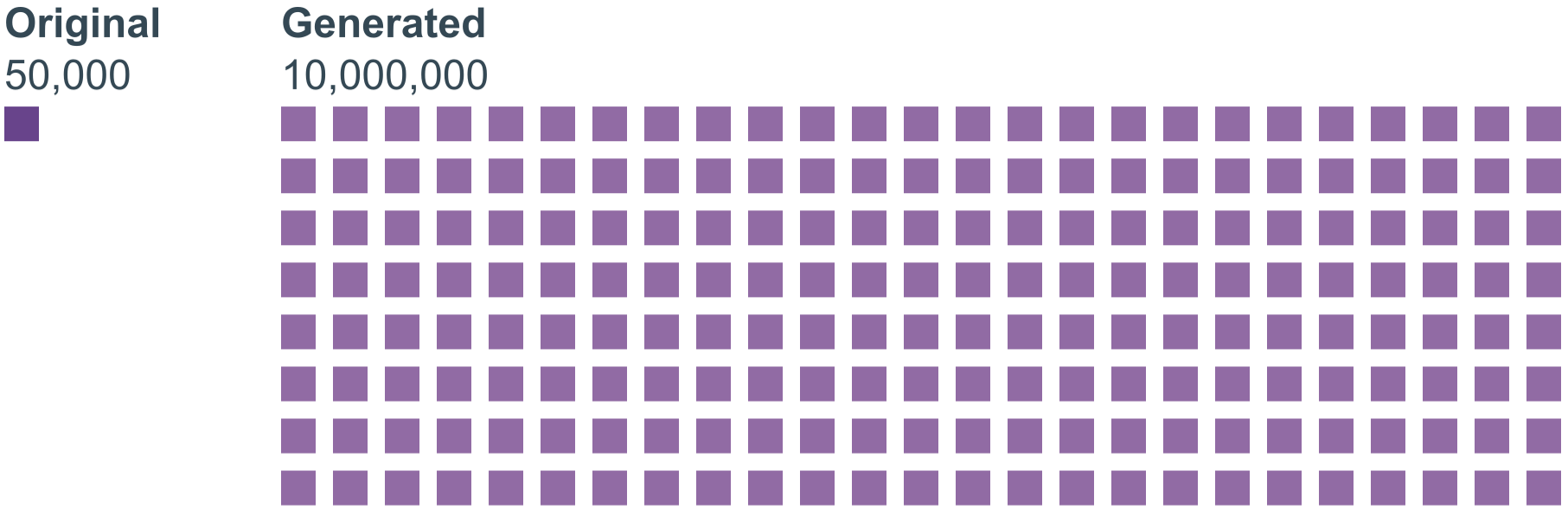
On the Scalability of Certified Adversarial Robustness with Generated Data

Thomas Altstidl, David Dobre, Arthur Kosmala, Björn Eskofier, Gauthier Gidel, Leo Schwinn

Empirical Robustness

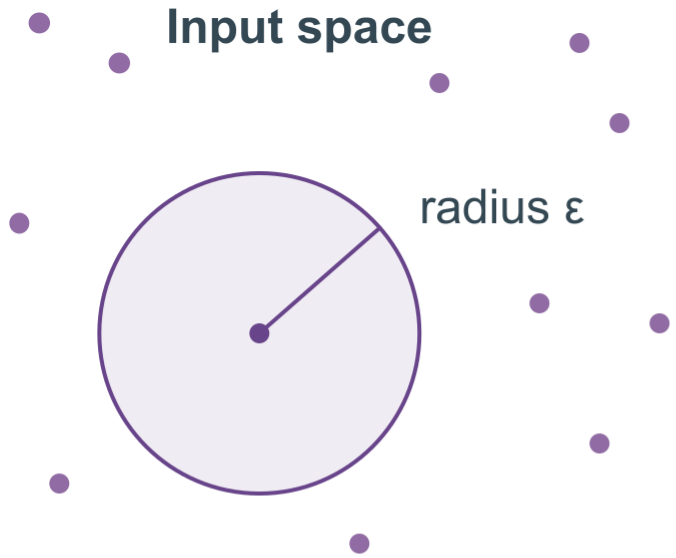


54.37%
AutoAttack accuracy with only original data

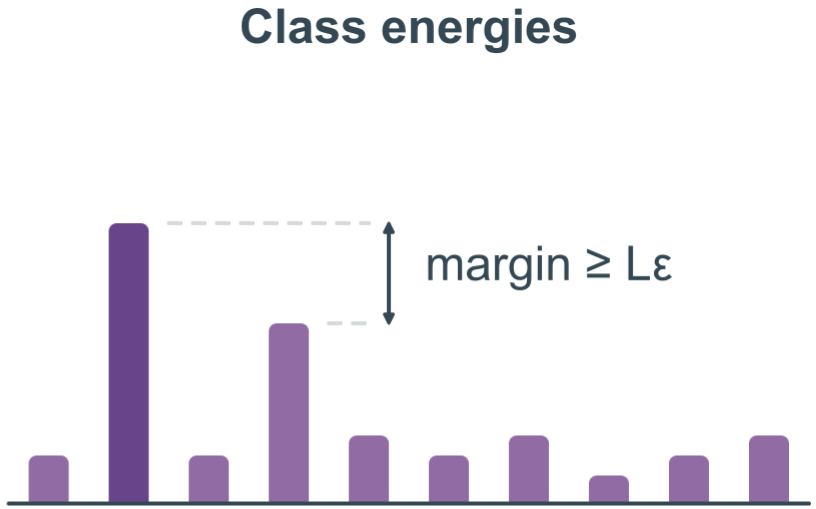
→

64.19%
AutoAttack accuracy with both original and generated data

Certified Robustness



Lipschitz constant L



What we show

39.72%

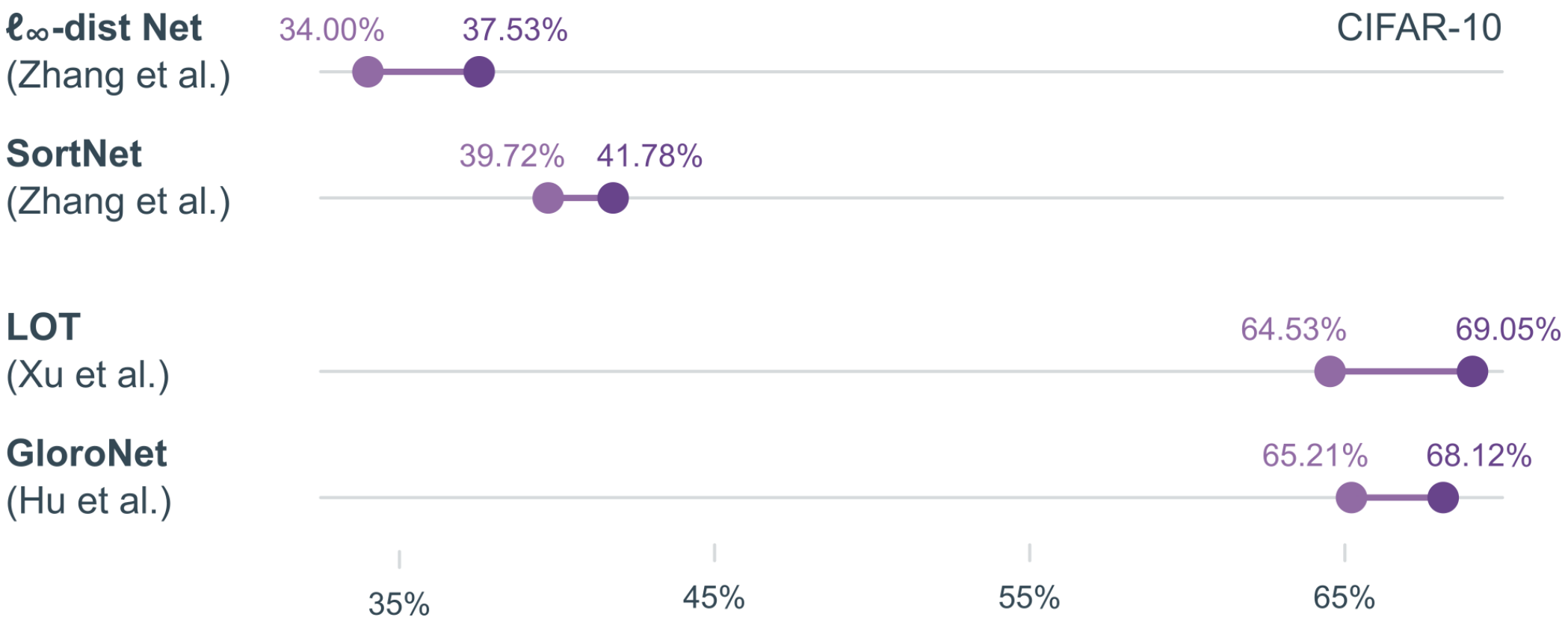
Certified accuracy with only original data

→

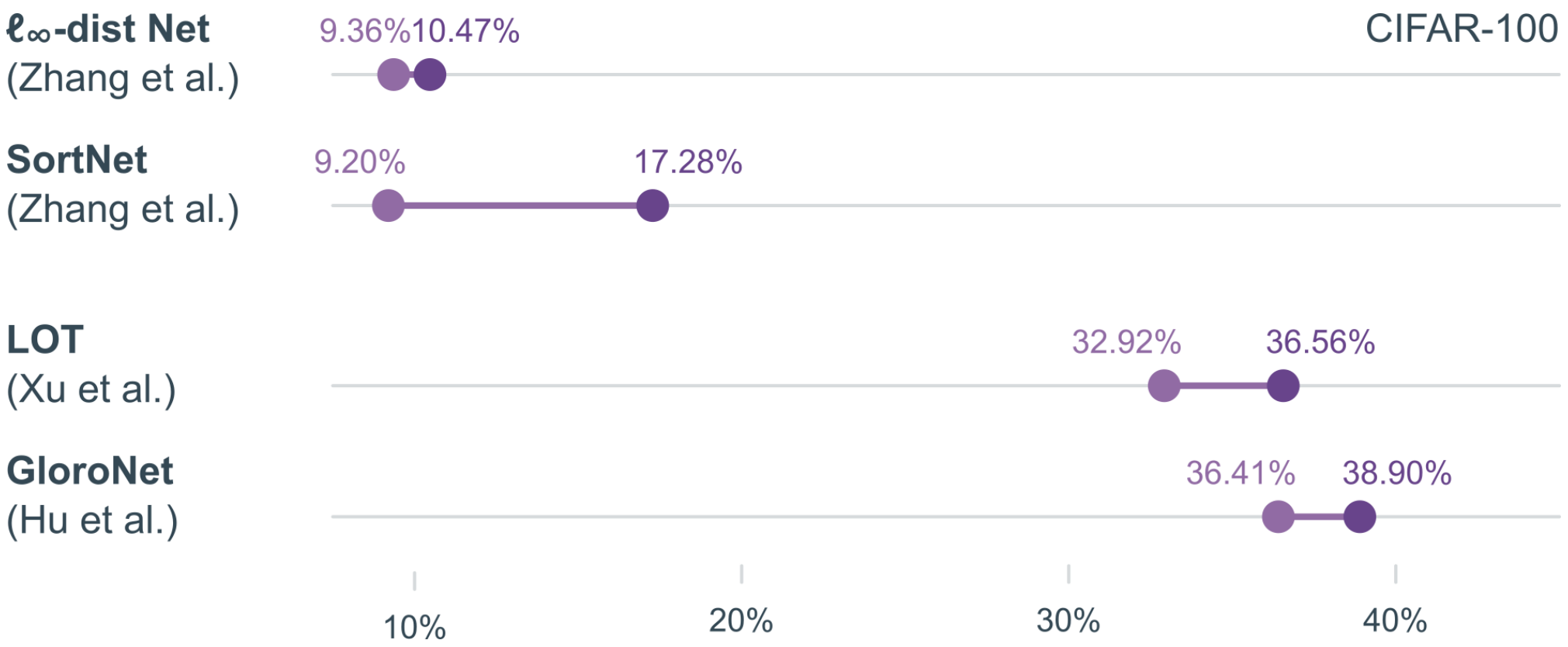
41.78%

Certified accuracy with both original and generated data

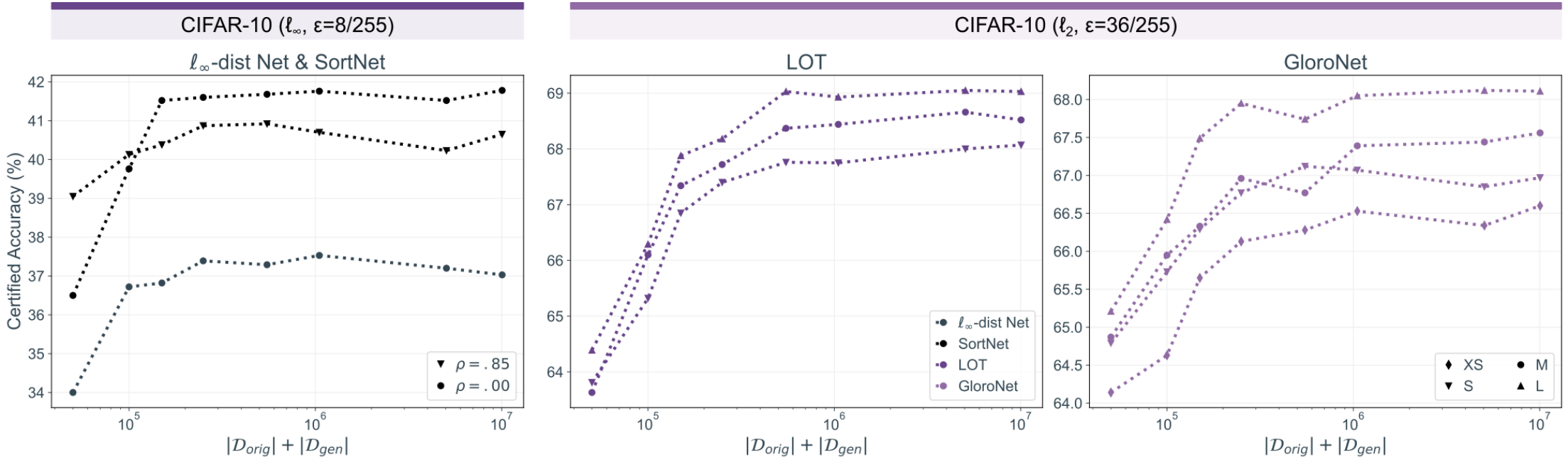
Experiments on CIFAR-10



Experiments on CIFAR-100



Scalability with Generated Data



Limits of Scalability

There seem to be clear limits to the scalability of certified adversarial robustness, with no substantial improvements after using 1 million generated CIFAR-10 images