A Little Robustness Goes a Long Way: Leveraging Robust Features for Targeted Transfer Attacks

Jacob M. Springer[1], **Melanie Mitchell**[2], **Garrett T. Kenyon**[1]

[1]Los Alamos National Laboratory [2]Santa Fe Institute

NeurIPS 2021

- Robust neural networks can increase targeted adversarial transferability

- Robust neural networks can increase targeted adversarial transferability
- Image representations are often highly transferable when generated with robust neural networks

- Why are adversarial examples optimized to fool robust neural networks so transferable?

- Why are adversarial examples optimized to fool robust neural networks so transferable?
- Why do non-robust networks not have this property?

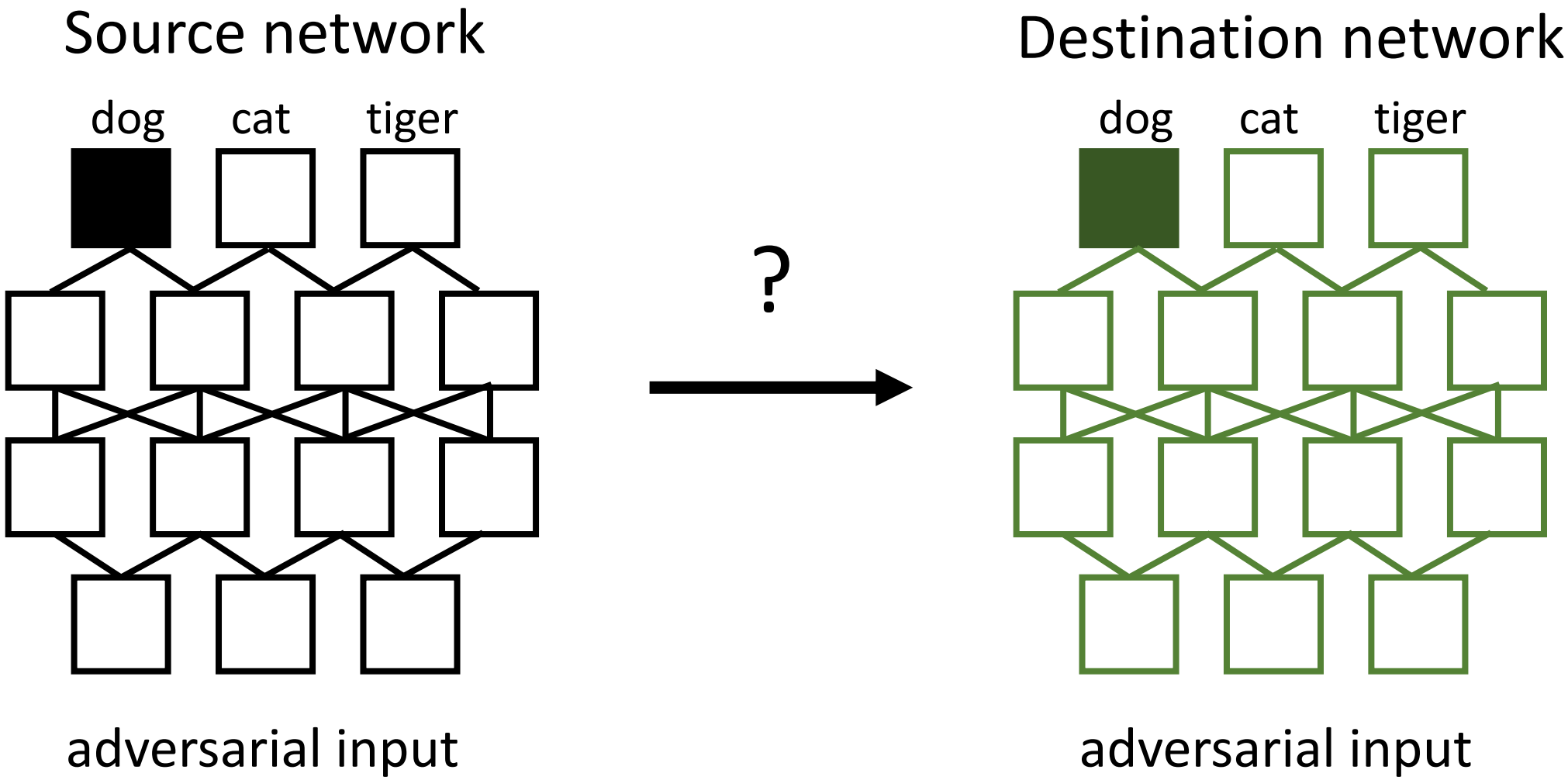
- Non-robust neural networks have low *representation transferability*

- Non-robust neural networks have low *representation transferability*
- Robust neural networks have high *representation transferability*

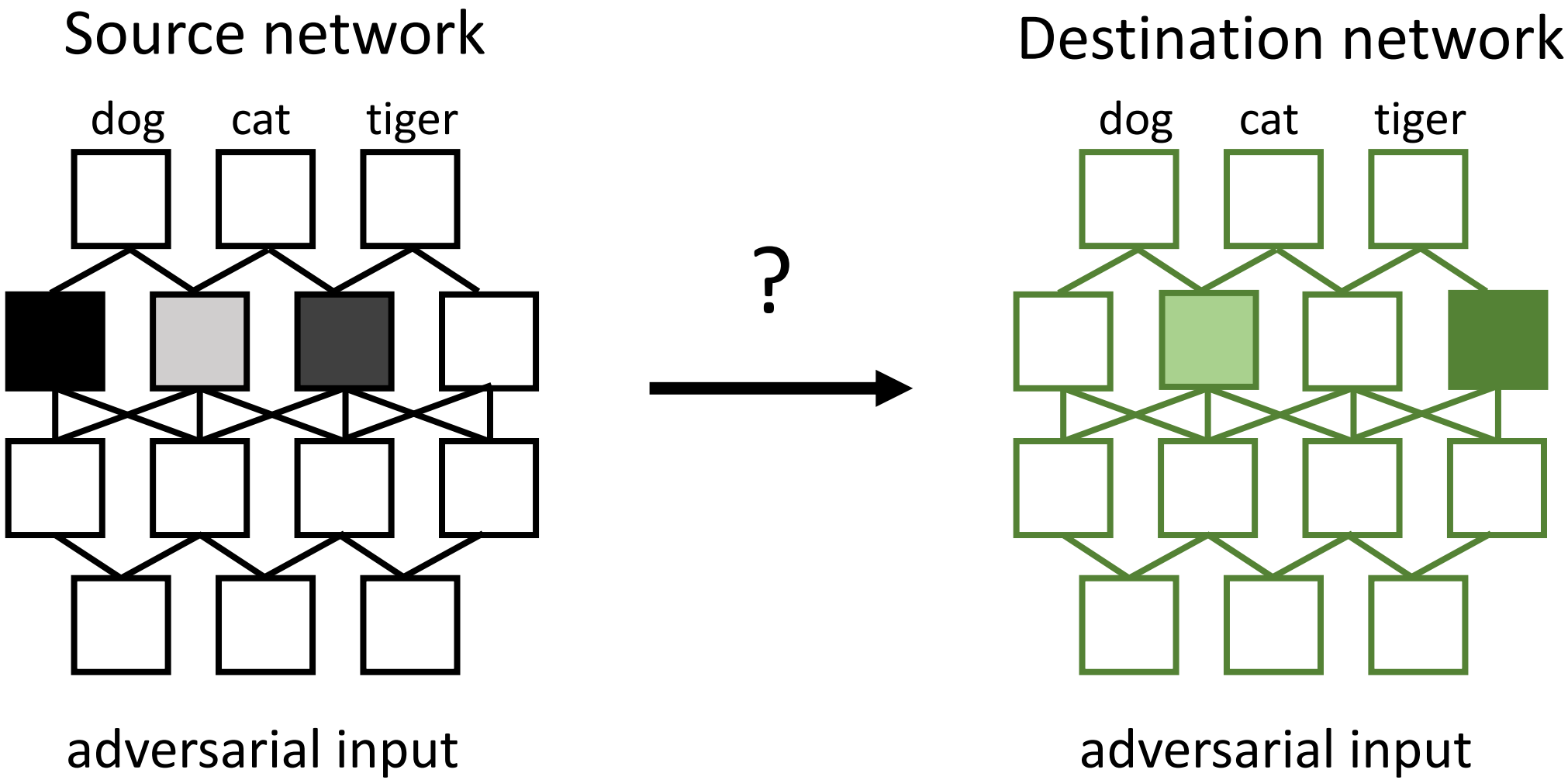
- *Robust* neural networks: adversarial training

- *Robust* neural networks: adversarial training
- Train on adversarial examples limited in L2 norm by ε (*robustness parameter)*

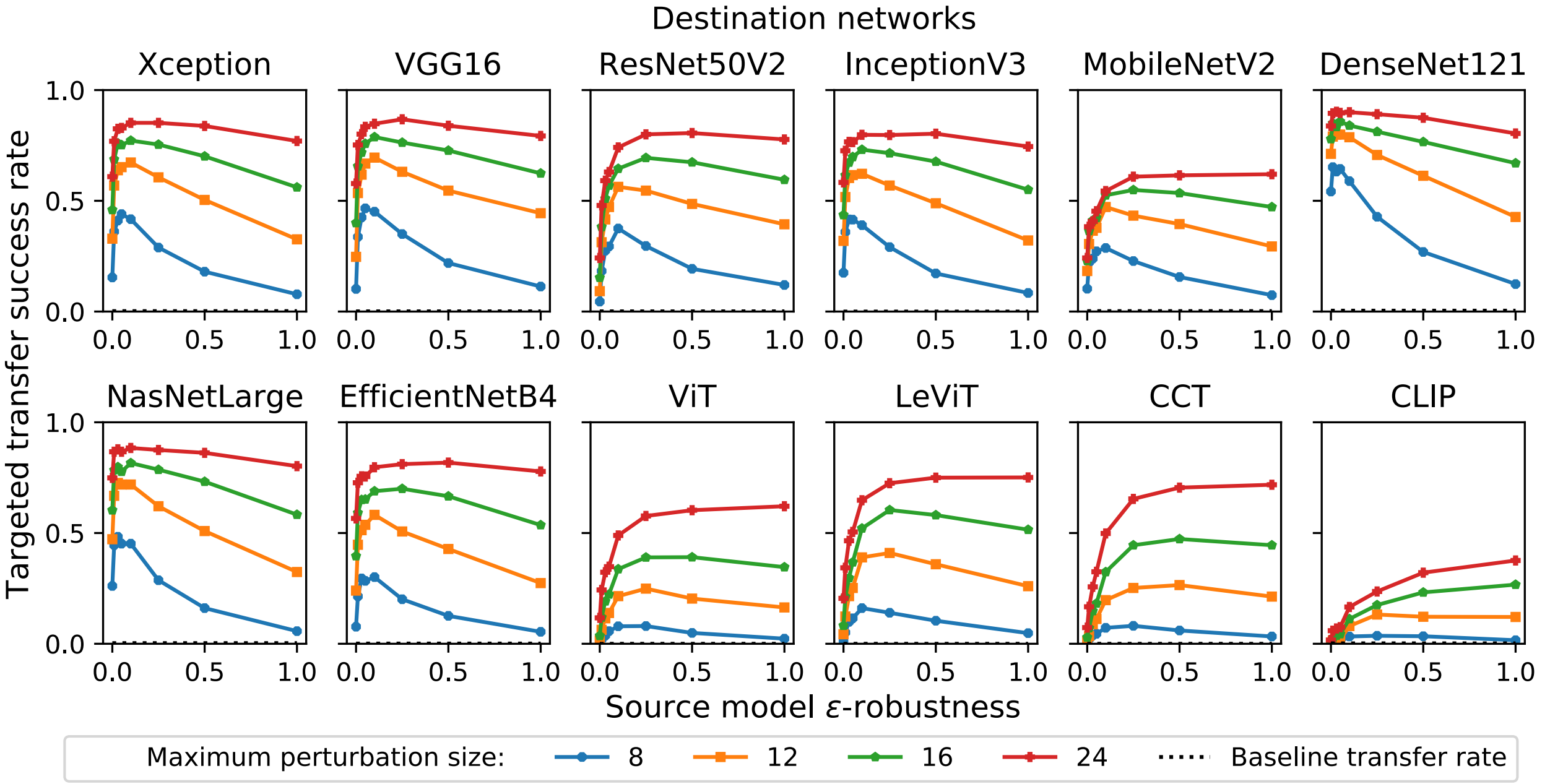
Class-targeted (standard) transferability

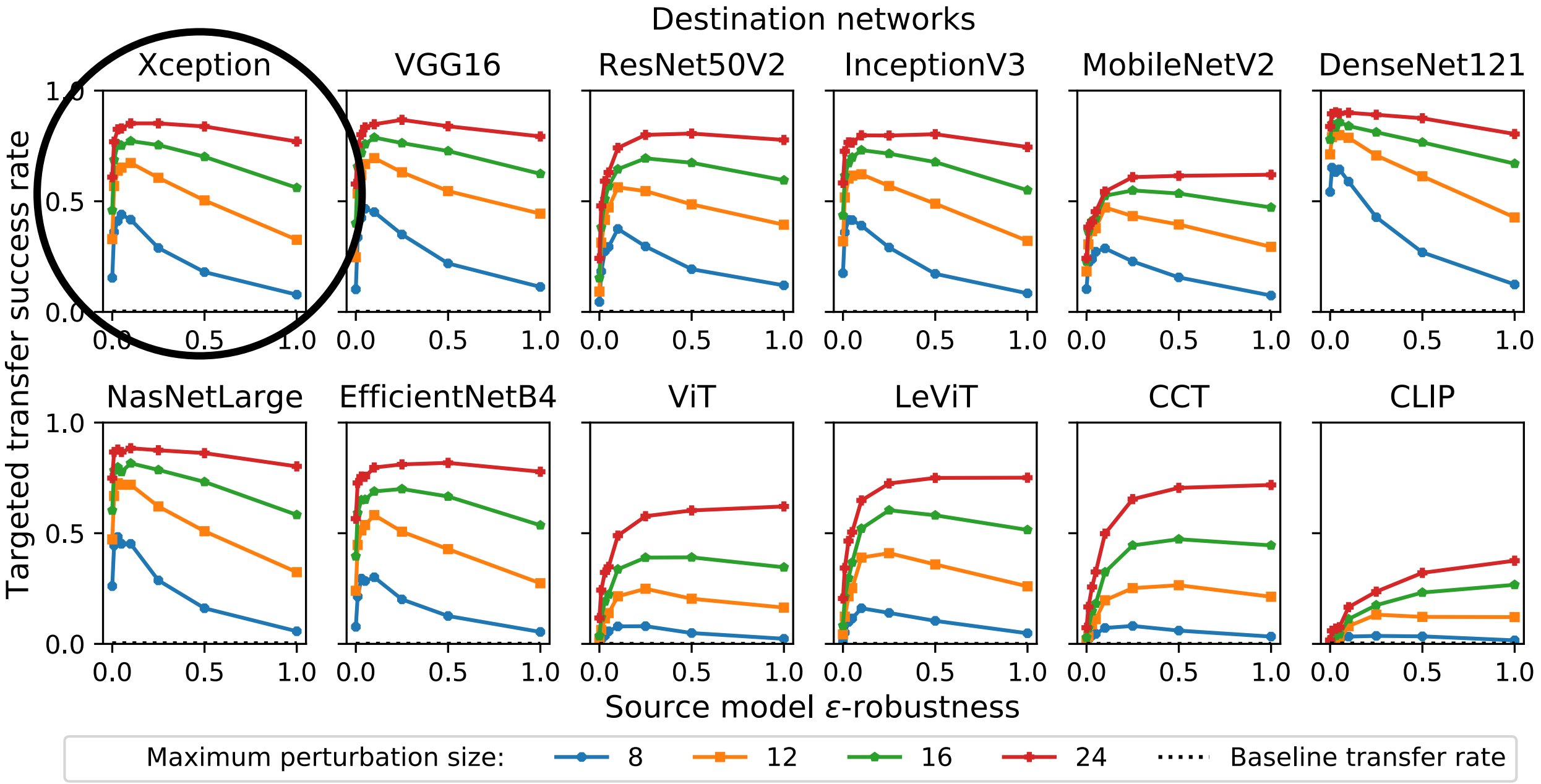


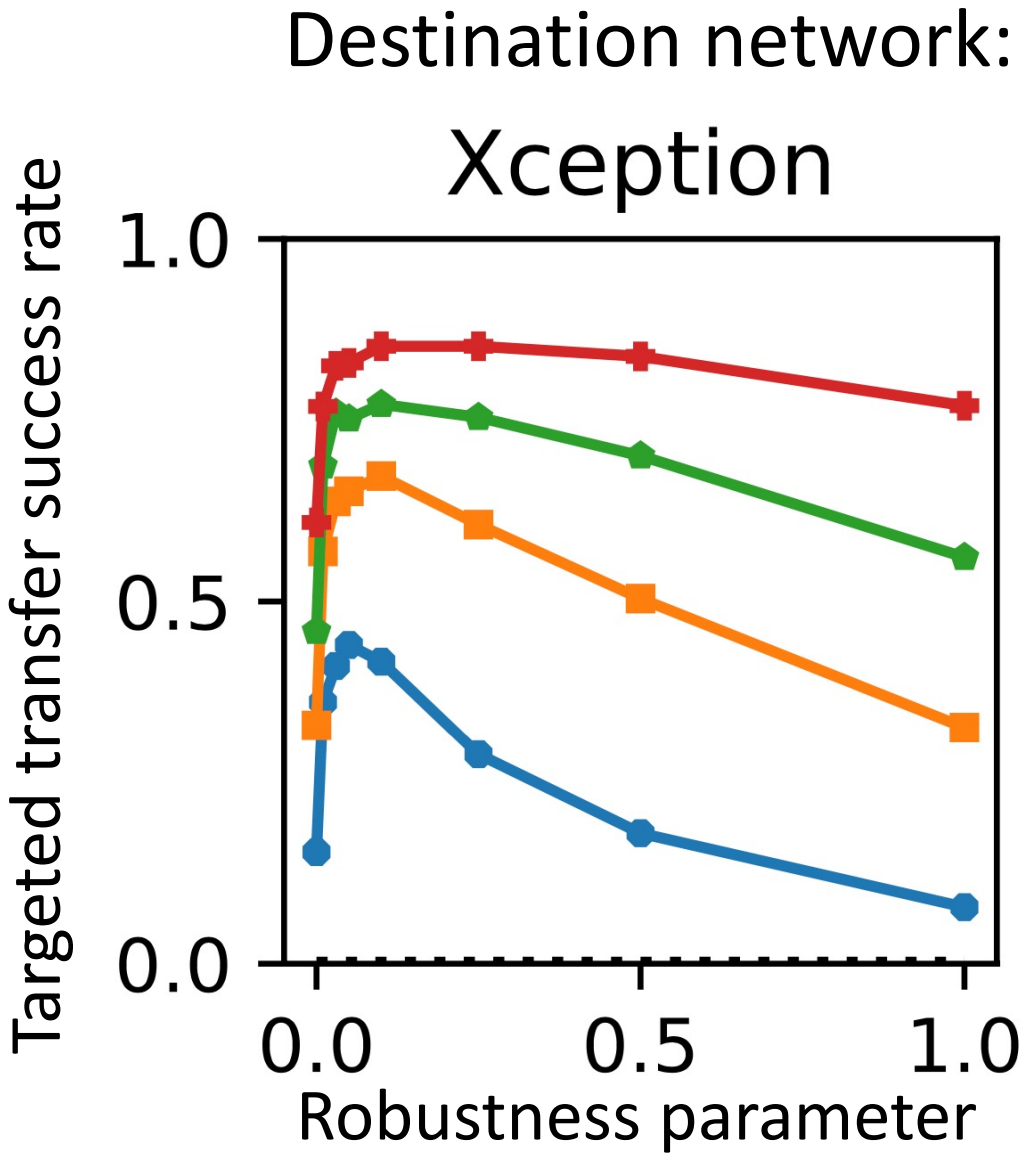
Representation-targeted transferability

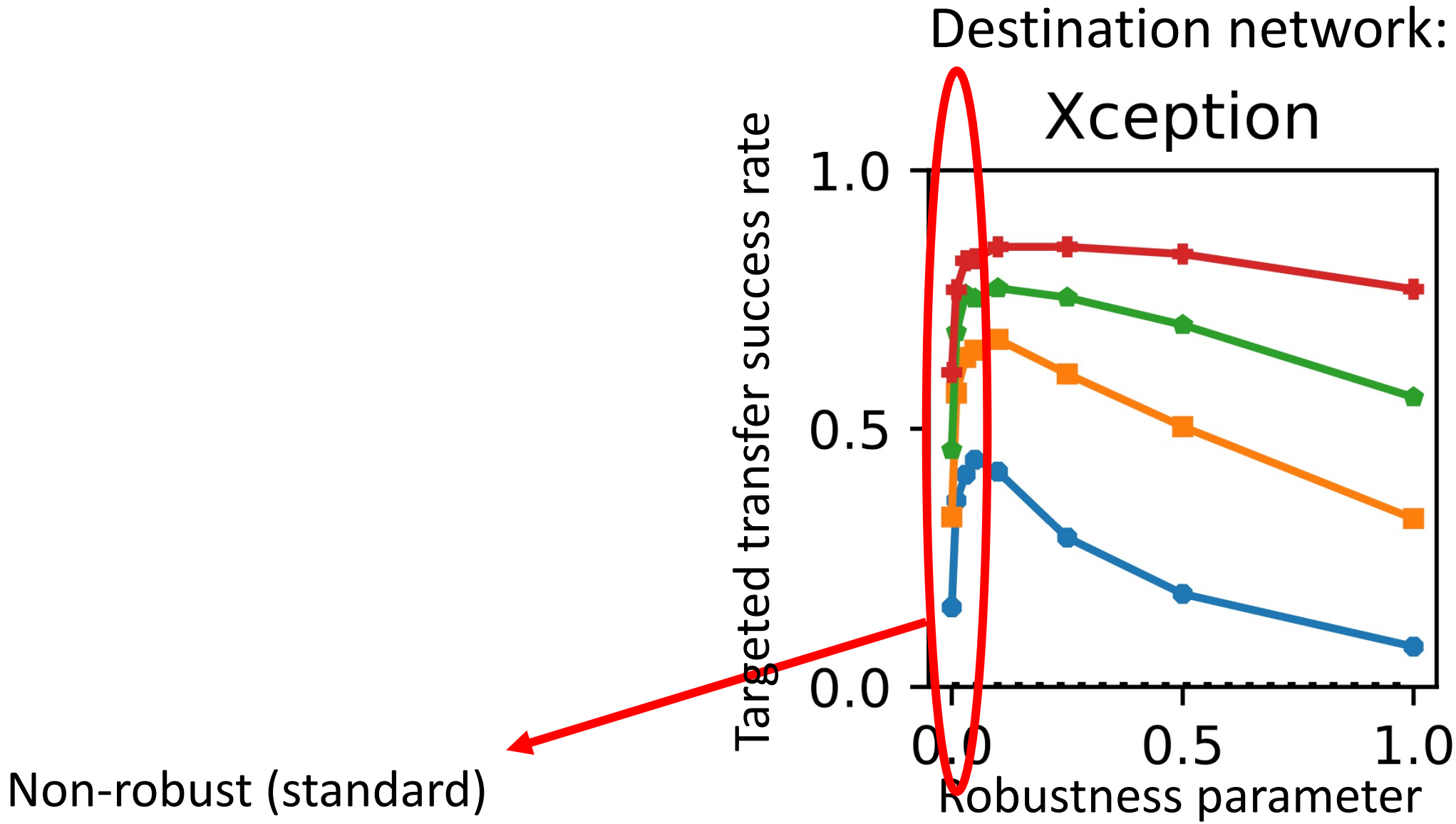


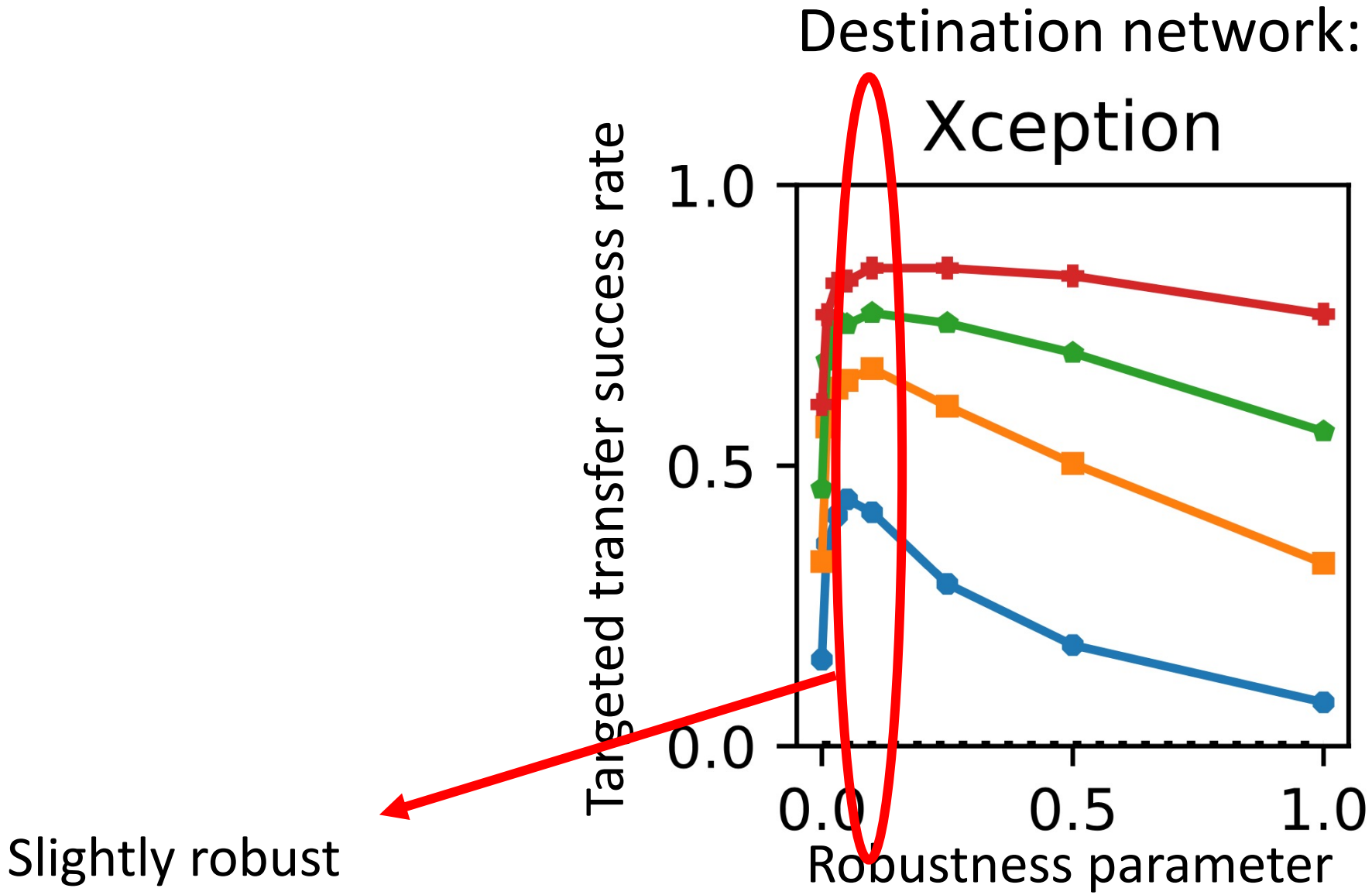
- Transferability in this talk is *targeted*
- See paper for experiments on *untargeted* adversarial examples

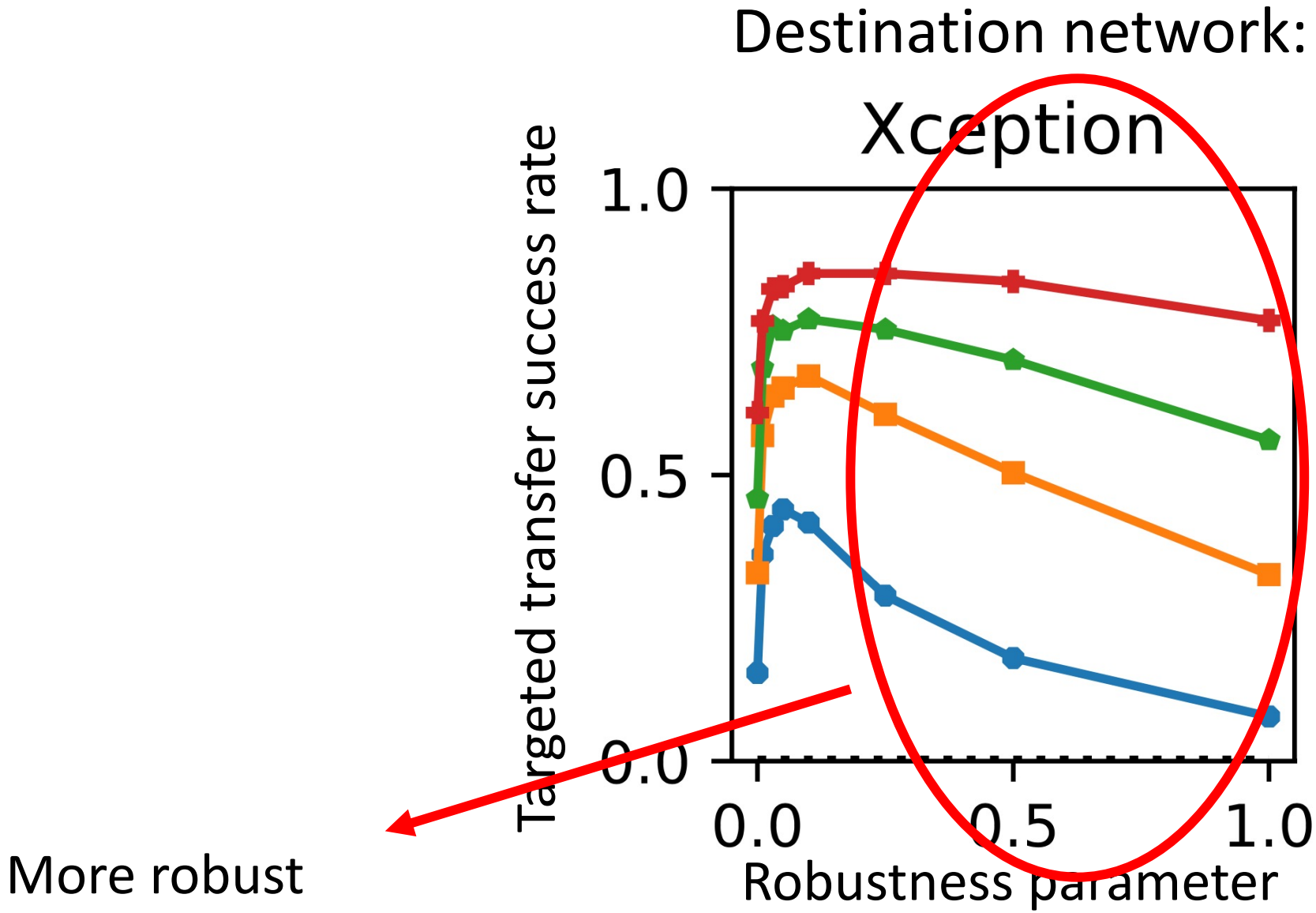

Destination networks
Xception
VGG16
ResNet50V2
InceptionV3
MobileNetV2
DenseNet121
NasNetLarge
EfficientNetB4
ViT
LeViT
CCT
CLIP
Targeted transfer success rate
Source model ε-robustness
0.0
0.5
1.0
Maximum perturbation size:
8
12
16
24
Baseline transfer rate

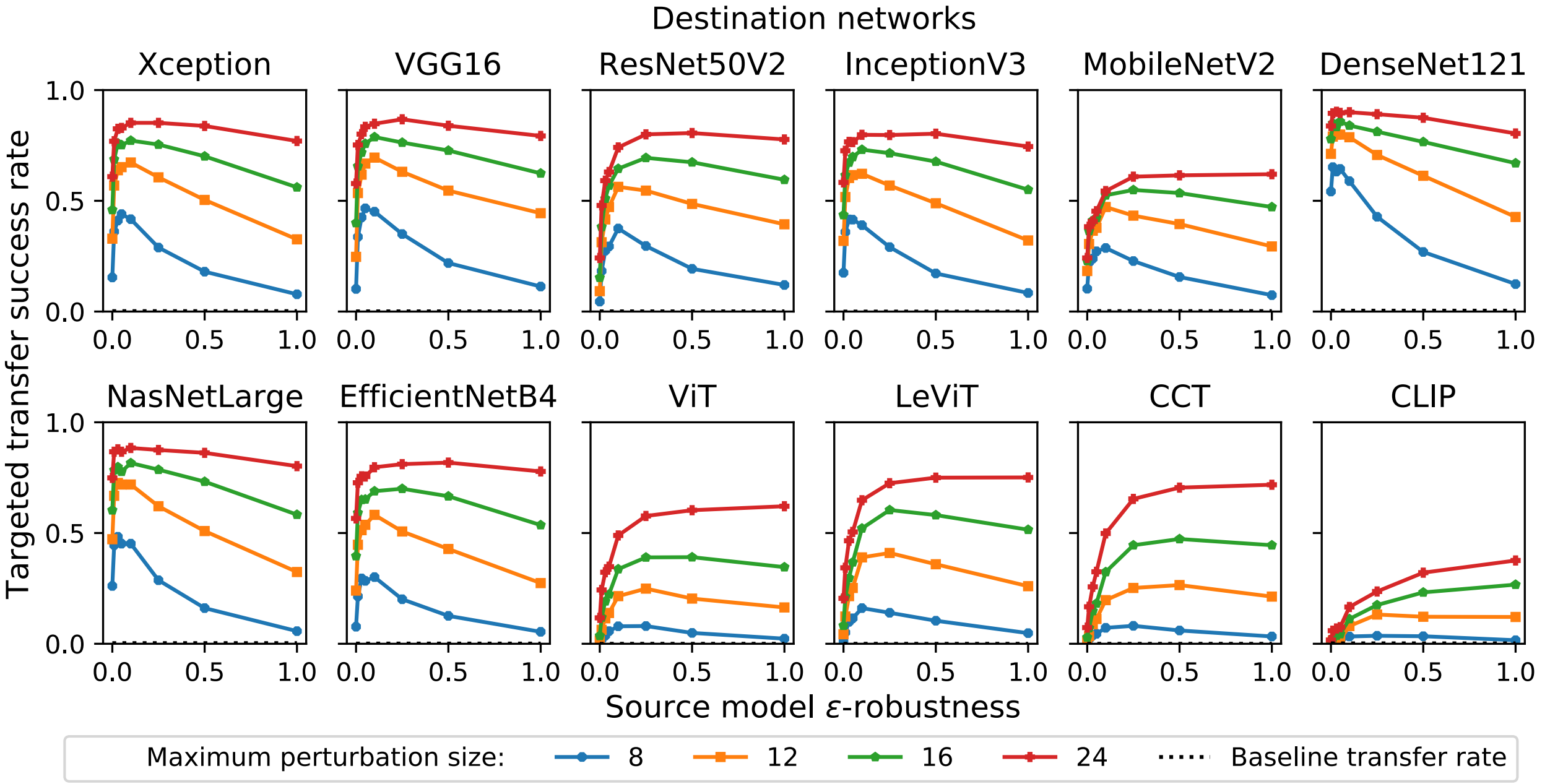

Destination networks
Xception
VGG16
ResNet50V2
InceptionV3
MobileNetV2
DenseNet121
NasNetLarge
EfficientNetB4
ViT
LeViT
CCT
CLIP
Targeted transfer success rate
Source model ε-robustness
0.0
0.5
1.0
Maximum perturbation size:
8
12
16
24
Baseline transfer rate

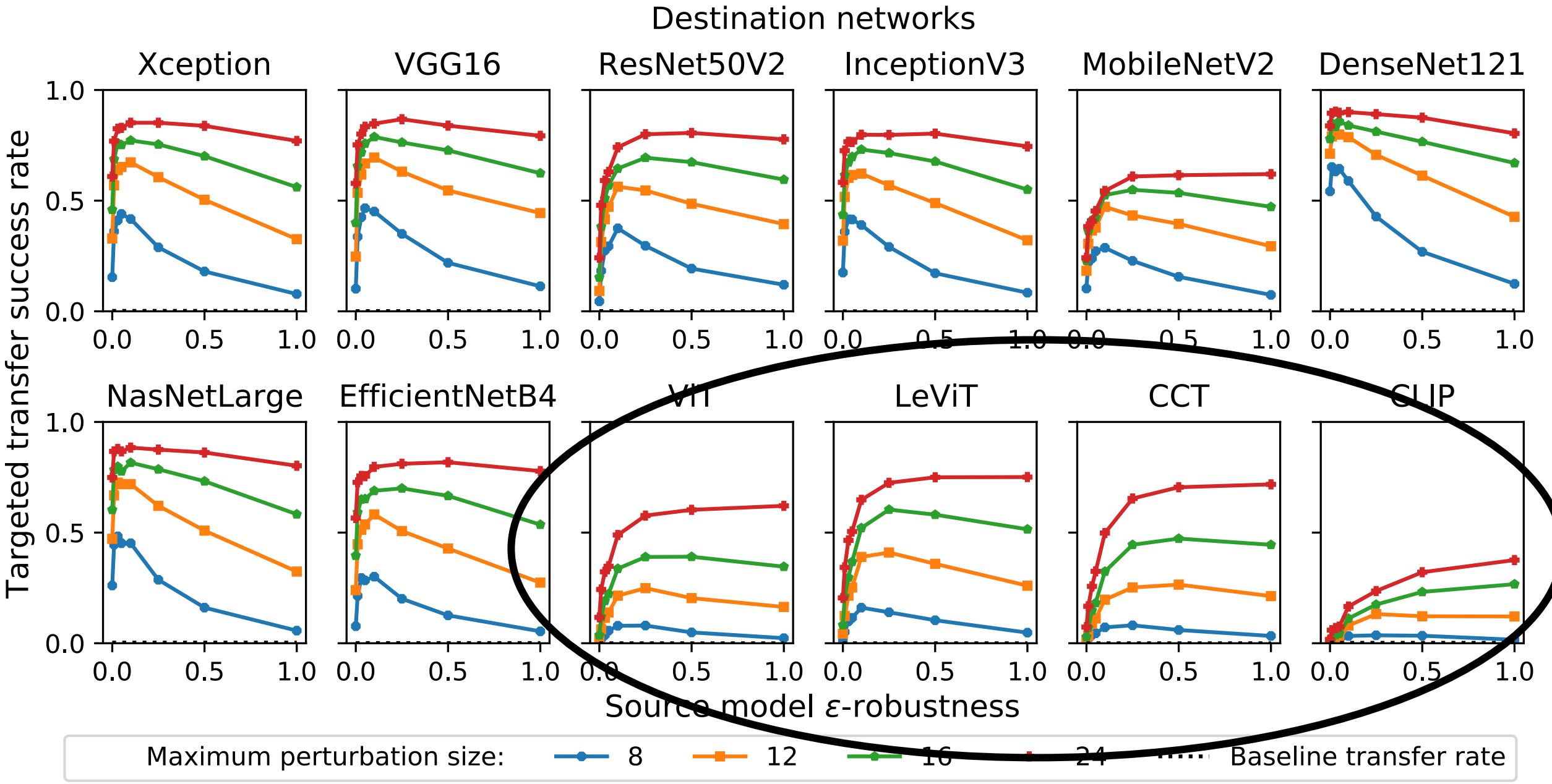

Destination network:
Xception
Targeted transfer success rate
1.0
0.5
0.0
0.0
0.5
1.0
Robustness parameter

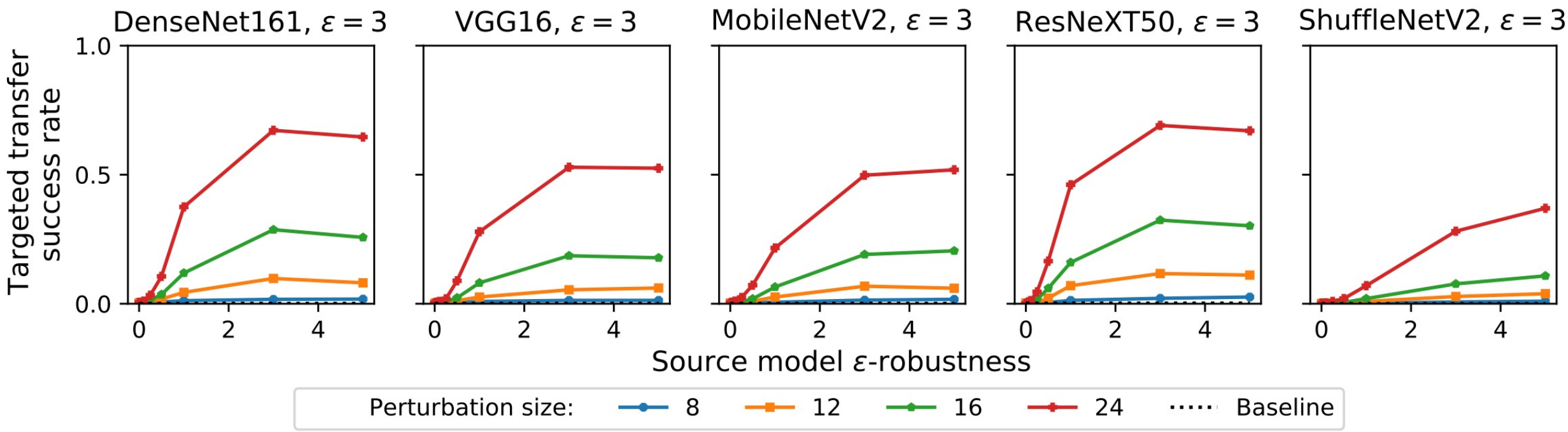

Destination network:
Xception
Targeted transfer success rate
1.0
0.5
0.0
0.0
0.5
1.0
Robustness parameter
Non-robust (standard)


Destination network:
Xception
Targeted transfer success rate
1.0
0.5
0.0
0.0
0.5
1.0
Robustness parameter
Slightly robust

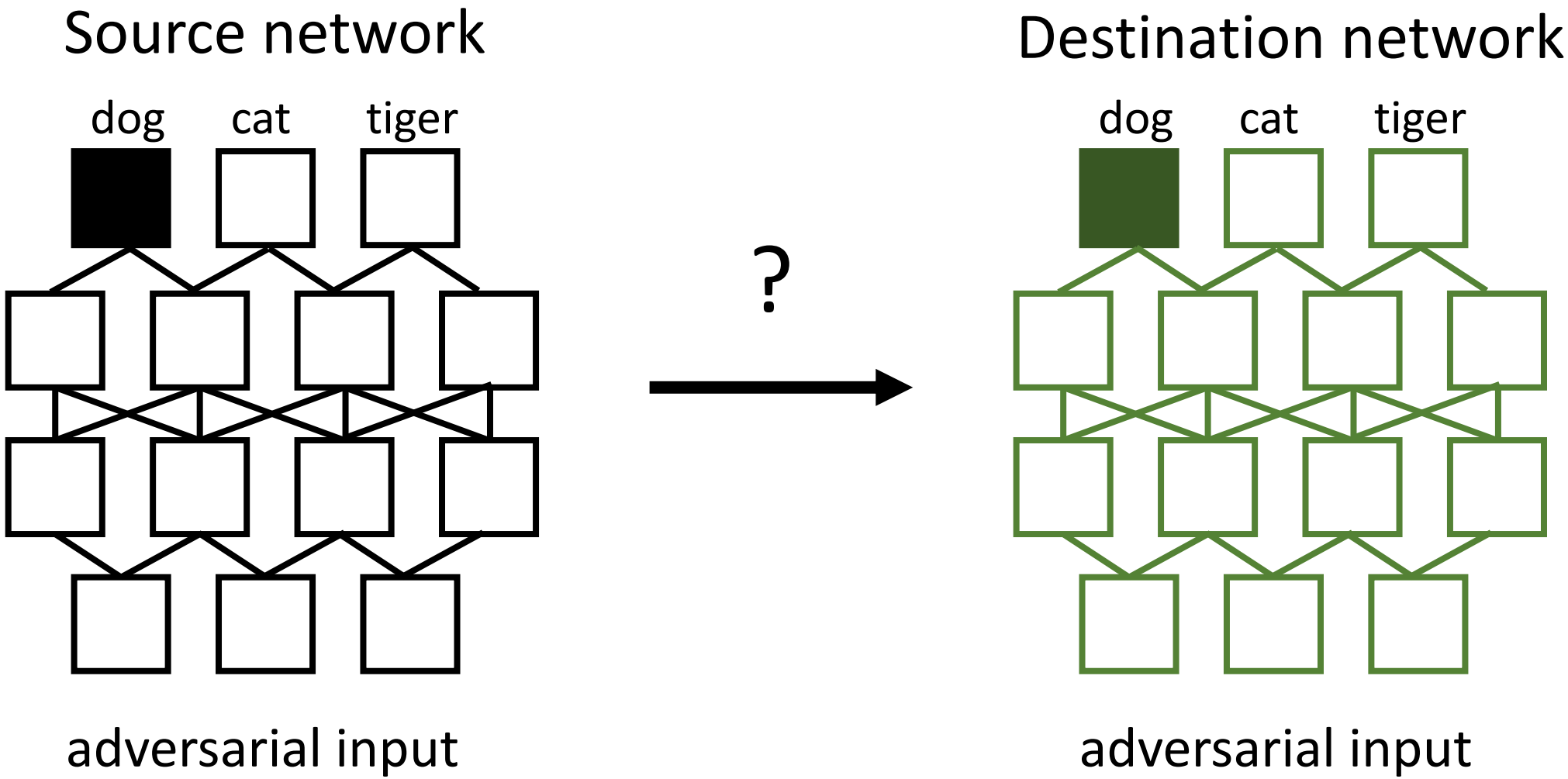

Destination network:
Xception
Targeted transfer success rate
1.0
0.5
0.0
0.0
0.5
1.0
Robustness parameter
More robust


Destination networks
Xception
VGG16
ResNet50V2
InceptionV3
MobileNetV2
DenseNet121
NasNetLarge
EfficientNetB4
ViT
LeViT
CCT
CLIP
Targeted transfer success rate
Source model ε-robustness
0.0
0.5
1.0
Maximum perturbation size:
8
12
16
24
Baseline transfer rate

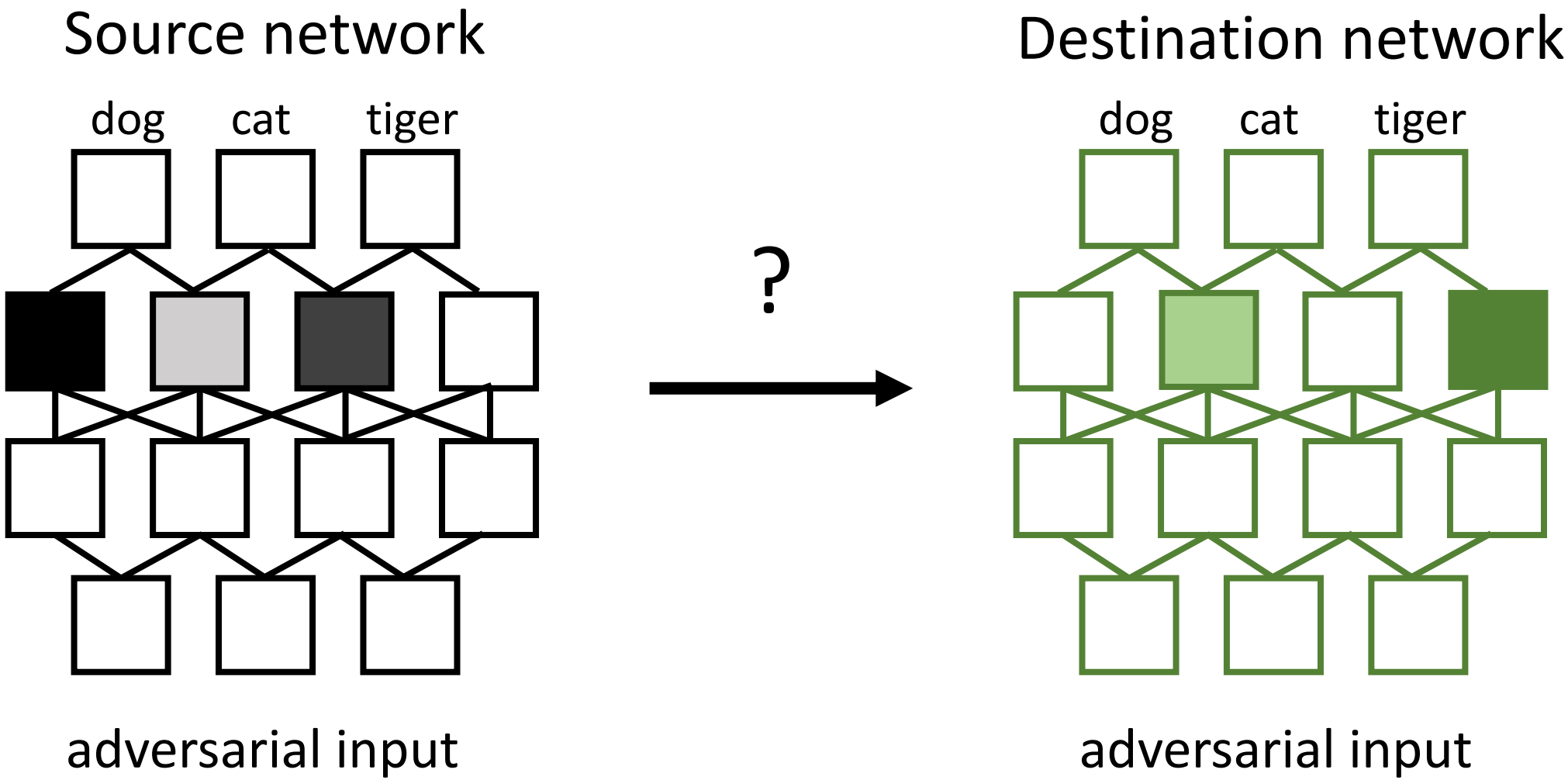

Destination networks
Xception
VGG16
ResNet50V2
InceptionV3
MobileNetV2
DenseNet121
NasNetLarge
EfficientNetB4
ViT
LeViT
CCT
CLIP
Targeted transfer success rate
Source model ε-robustness
0.0
0.5
1.0
Maximum perturbation size:
8
12
16
24
Baseline transfer rate


DenseNet161, $\varepsilon = 3$
VGG16, $\varepsilon = 3$
MobileNetV2, $\varepsilon = 3$
ResNeXT50, $\varepsilon = 3$
ShuffleNetV2, $\varepsilon = 3$
Targeted transfer success rate
1.0
0.5
0.0
0
2
4
Source model ε-robustness
Perturbation size:
8
12
16
24
Baseline

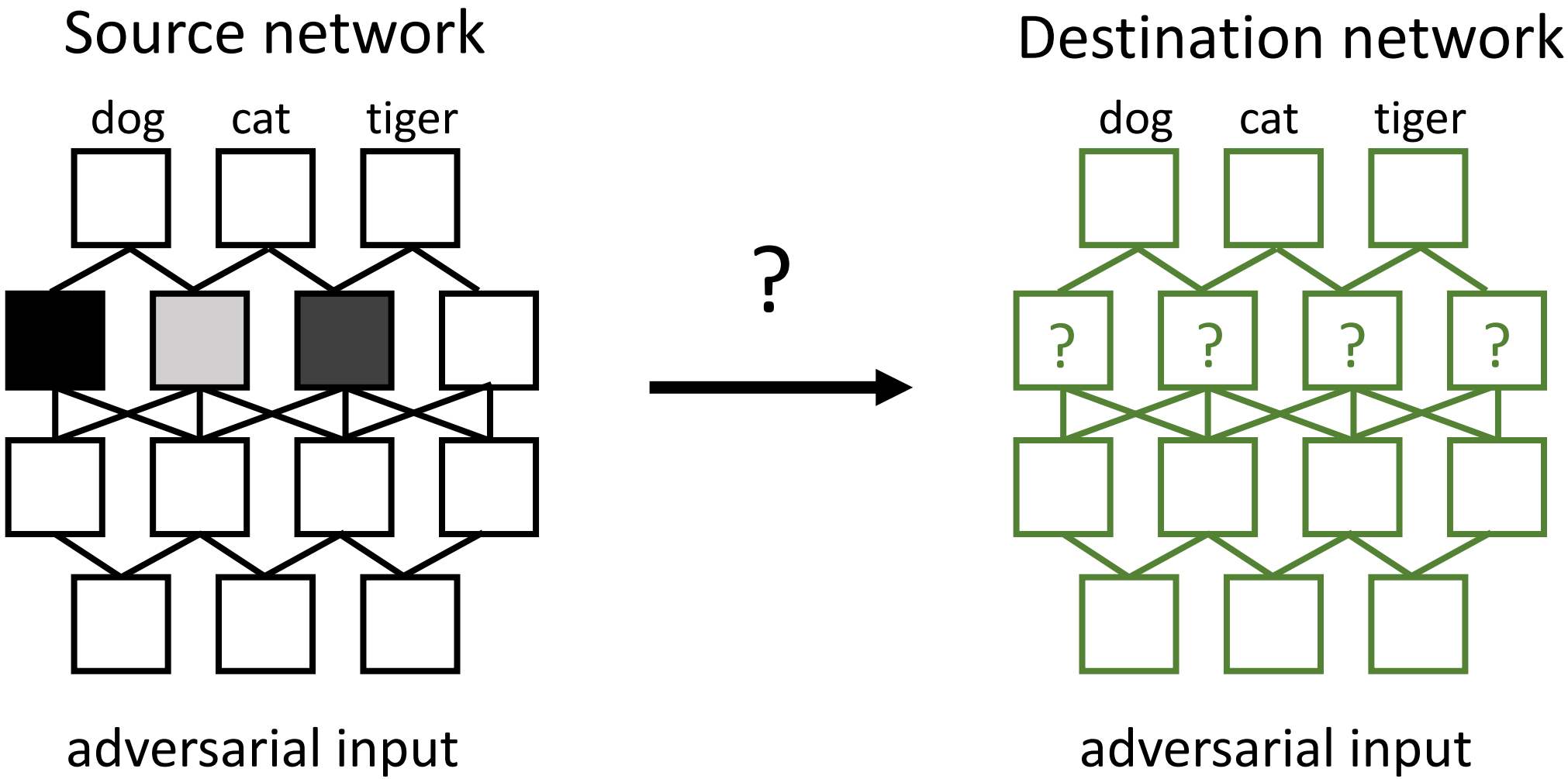
Class-targeted transferability



Representation-targeted transferability



Representation-targeted transferability



Source classifier

Destination classifier

reference

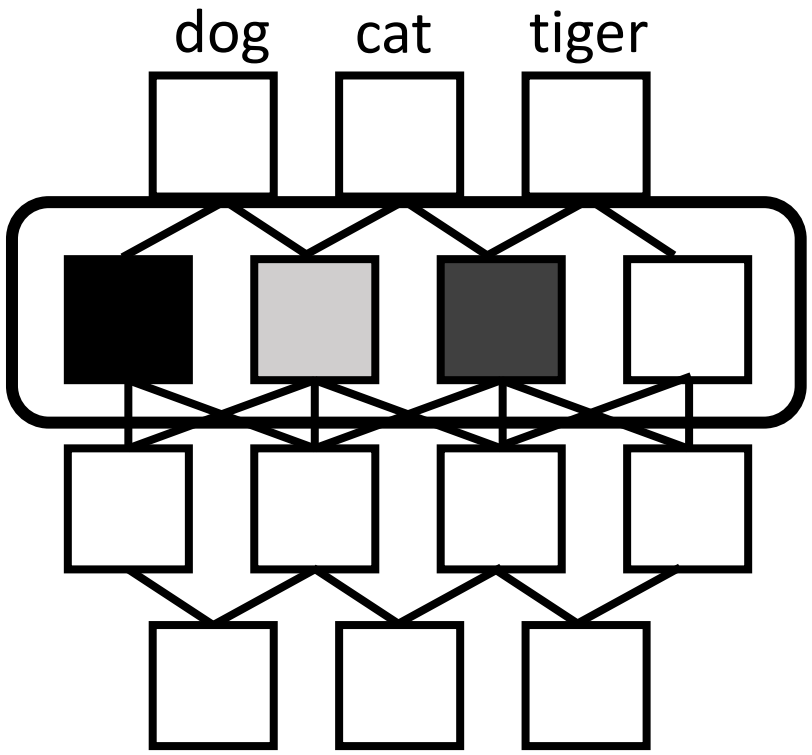
dog cat tiger

reference in destination

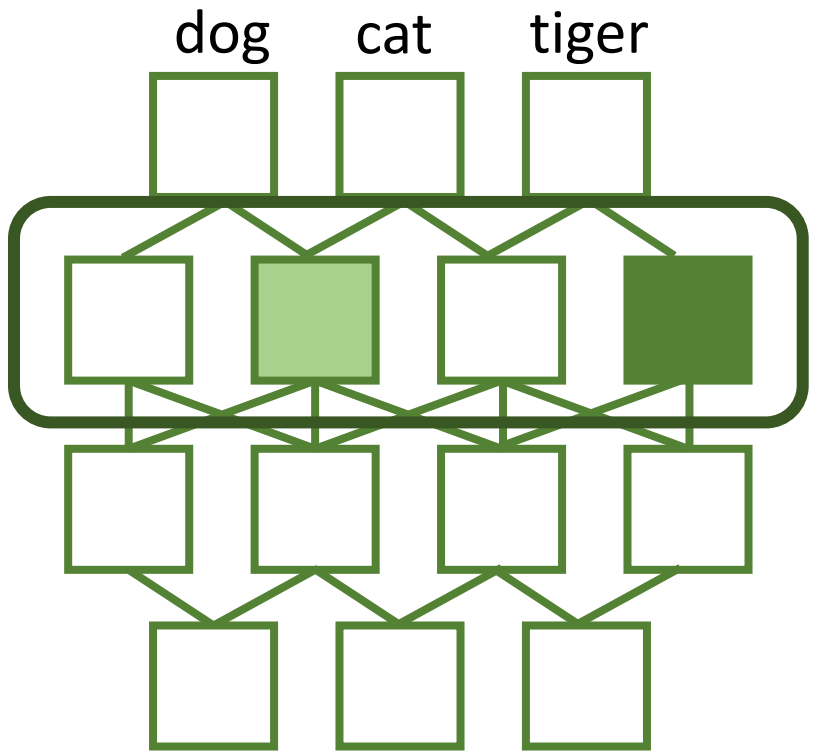
dog cat tiger

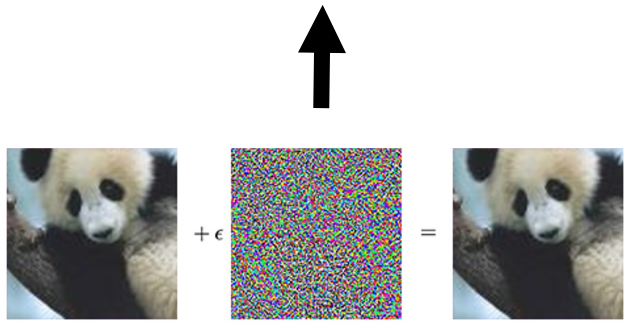
Source classifier

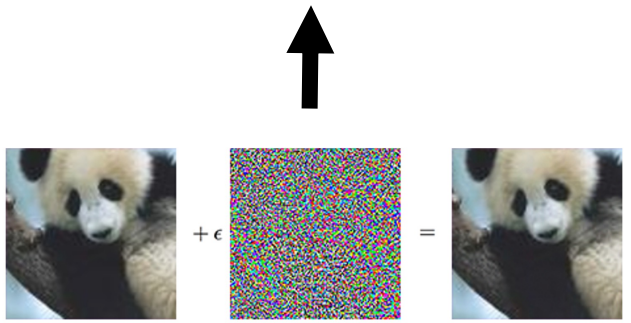
Destination classifier

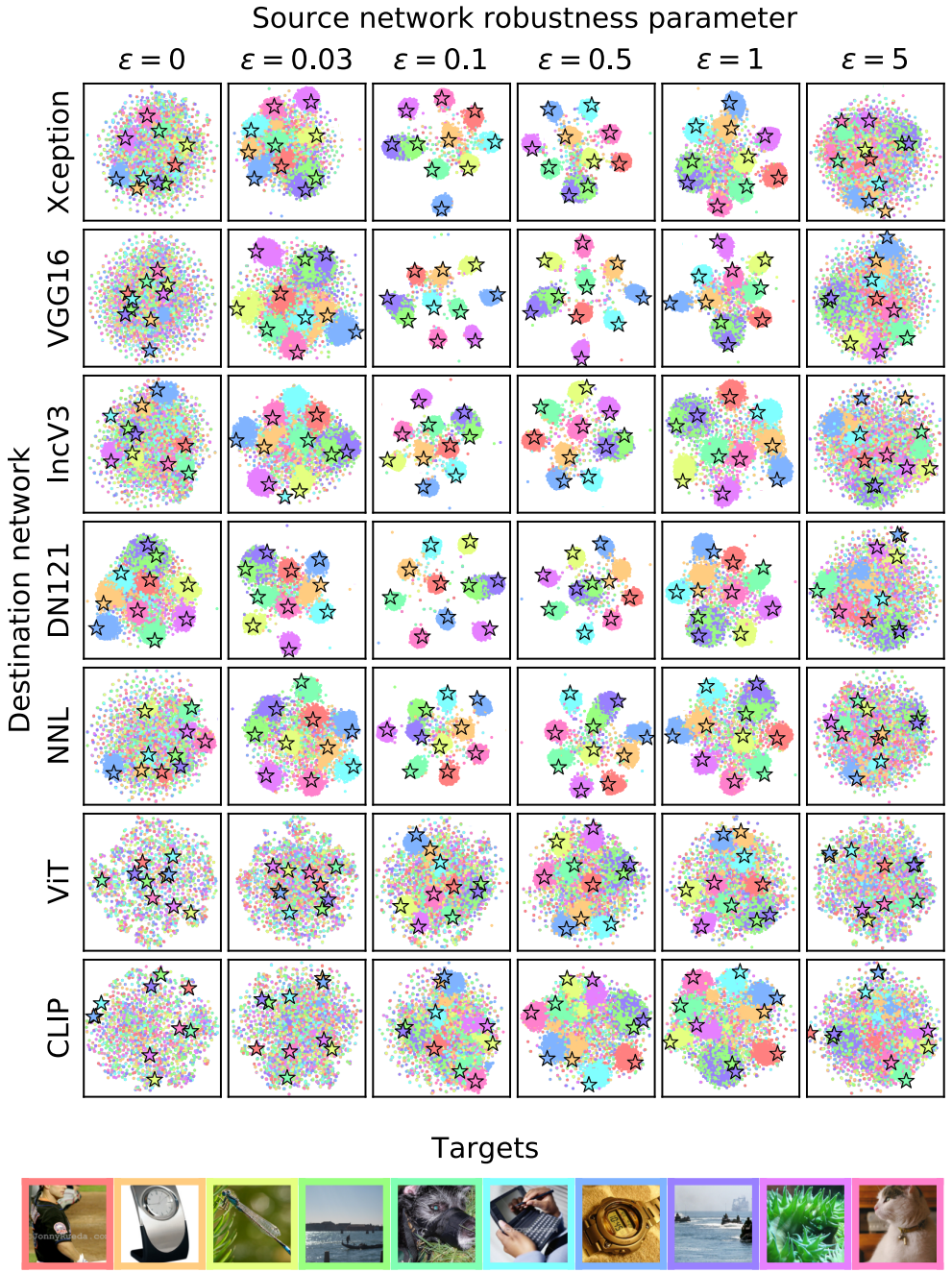


?








Source network robustness parameter
$\varepsilon = 0$
$\varepsilon = 0.03$
$\varepsilon = 0.1$
$\varepsilon = 0.5$
$\varepsilon = 1$
$\varepsilon = 5$
Destination network
Xception
VGG16
IncV3
DN121
NNL
ViT
CLIP
Targets

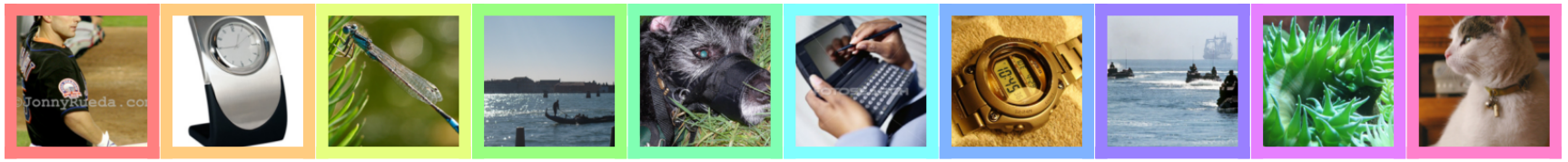
Source network robustness parameter

$\varepsilon = 0$ $\varepsilon = 0.03$ $\varepsilon = 0.1$ $\varepsilon = 0.5$ $\varepsilon = 1$ $\varepsilon = 5$

Xception

CLIP

Targets



Destination	Source network robustness parameter (ε)									
	0	0.01	0.03	0.05	0.1	0.25	0.5	1	3	5
Xception	0.462	0.505	0.531	0.563	**0.594**	0.585	0.572	0.543	0.449	0.404
VGG16	0.333	0.401	0.417	0.494	**0.528**	0.520	0.520	0.486	0.383	0.333
ResNet50V2	0.284	0.348	0.379	0.432	0.497	0.496	**0.510**	0.484	0.380	0.321
InceptionV3	0.577	0.612	0.627	0.644	**0.673**	0.662	0.655	0.636	0.572	0.539
MobileNetV2	0.431	0.459	0.460	0.493	**0.517**	0.513	0.513	0.504	0.455	0.425
DenseNet121	0.672	0.689	0.685	0.713	**0.726**	0.714	0.706	0.679	0.616	0.584
NasNetLarge	0.356	0.422	0.452	0.488	**0.541**	0.513	0.482	0.437	0.315	0.271
EfficientNetB4	0.085	0.111	0.137	0.144	**0.237**	0.220	0.226	0.202	0.112	0.074
ViT	0.066	0.087	0.109	0.129	0.195	**0.206**	**0.206**	0.203	0.120	0.086
CLIP	0.529	0.541	0.550	0.563	0.585	0.599	0.606	**0.613**	0.581	0.566

Cosine similarity of representation layer activations of adversarial example and target, in destination networks

1. Adversarial training of source networks can improve transferability of adversarial examples

2. Slightly-robust classifiers transfer features effectively, but non-robust classifiers do not