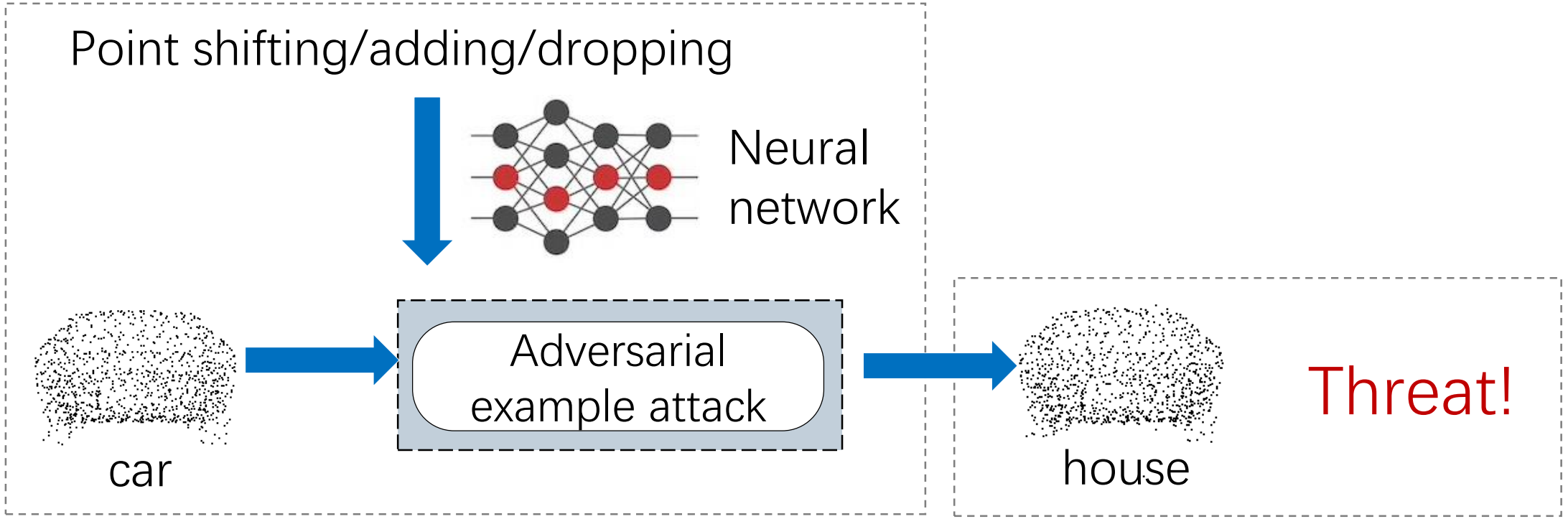
LG-GAN: Label Guided Adversarial Network for Flexible Targeted Attack of Point Cloud-based Deep Networks

Hang Zhou[1] Dongdong Chen[2] Jing Liao[3]
Kejiang Chen[1] Xiaoyi Dong[1] Kunlin Liu[1]
Weiming Zhang[1] Gang Hua[4] Nenghai Yu[1]

[1]University of Science and Technology of China [2]Microsoft Research
[3]City University of Hong Kong [4]Wormpex AI Research

Problem


Point shifting/adding/dropping
Neural network
Adversarial example attack
car
house
Threat!

Motivation

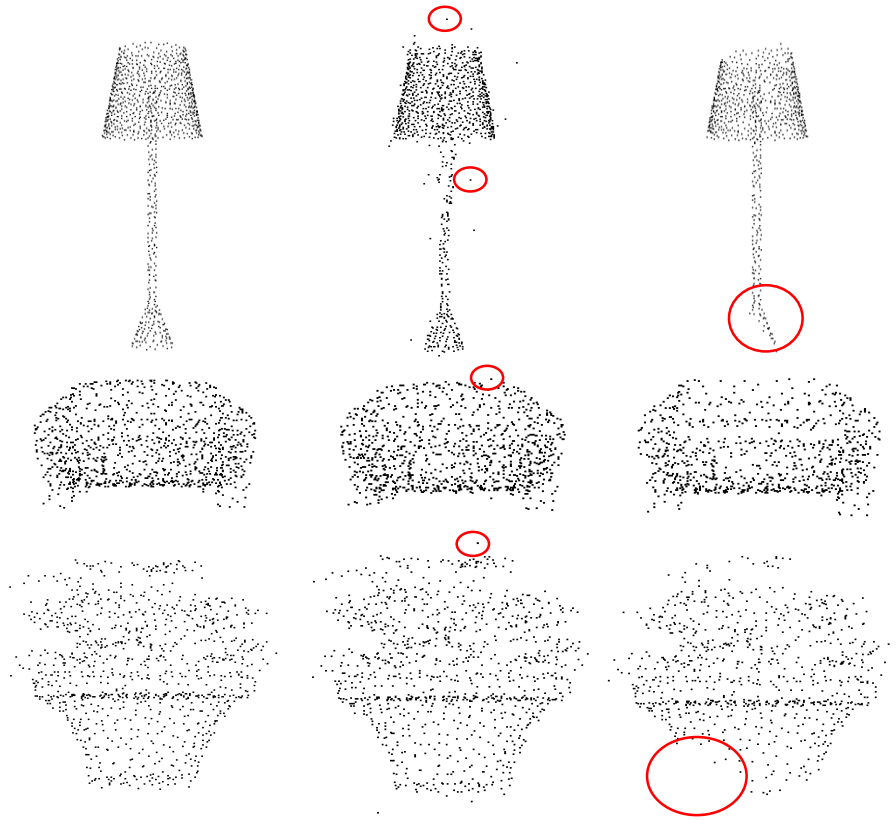
Related work

Current attack methods:

- Optimization-based:
  High attack success rate/**slow runtime**/**visible outliers**
- Gradient-based:
  Fast runtime/**low attack success rate**

Motivation

Generation based adversarial examples will avoid creating outliers and be fast in generation with high attack success rates.

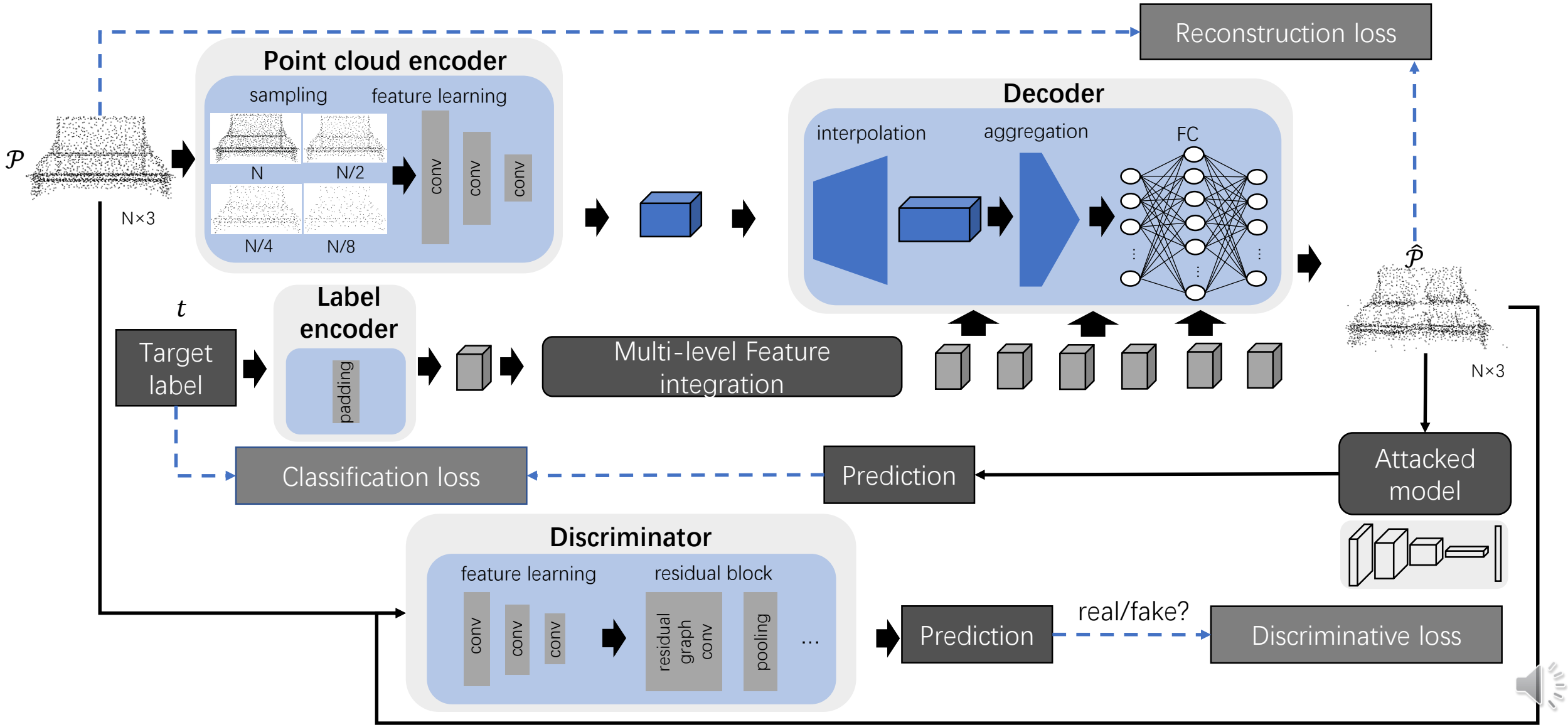


original point cloud

optimization based adversarial example

gradient based adversarial example

Framework


Reconstruction loss
Point cloud encoder
sampling
feature learning
N
N/2
N/4
N/8
conv
conv
conv
$\mathcal{P}$
N×3
Decoder
interpolation
aggregation
FC
$\hat{\mathcal{P}}$
N×3
t
Target label
Label encoder
padding
Multi-level Feature integration
Classification loss
Prediction
Attacked model
Discriminator
feature learning
residual block
conv
conv
conv
residual graph conv
pooling
...
Prediction
real/fake?
Discriminative loss

Objective loss functions

Generator:

$$\mathcal{L}_{\mathcal{G}} = \mathcal{L}_{cls} + \alpha\mathcal{L}_{rec} + \beta\mathcal{L}_{dis}$$

$$\mathcal{L}_{cls} = -\left[t \log \mathcal{H}\left(\hat{\mathcal{P}}\right) + (1-t) \log\left(\mathcal{H}\left(1-\hat{\mathcal{P}}\right)\right)\right]$$

where $\hat{\mathcal{P}} = \mathcal{G}_\theta(\mathcal{P}, t)$

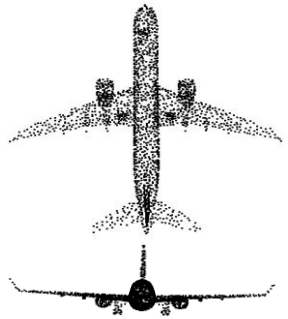
$\mathcal{L}_{rec}$ is ℓ_2 distance

$$\mathcal{L}_{dis}\left(\hat{\mathcal{P}}\right) = \left\|1 - D_\theta\left(\hat{\mathcal{P}}\right)\right\|_2^2$$

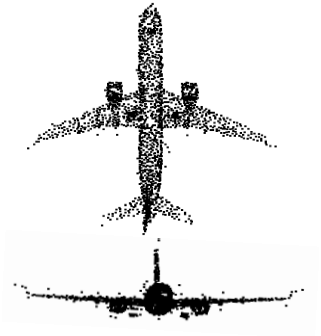
Discriminator:

$$\mathcal{L}_D\left(\mathcal{P}, \hat{\mathcal{P}}\right) = \frac{1}{2}\left\|D_\theta\left(\hat{\mathcal{P}}\right)\right\|_2^2 + \frac{1}{2}\left\|1 - D_\theta(\mathcal{P})\right\|_2^2$$

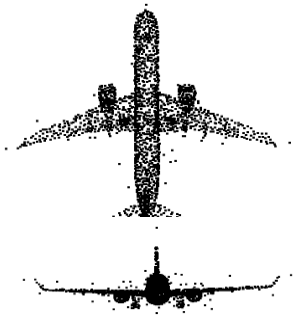
Results



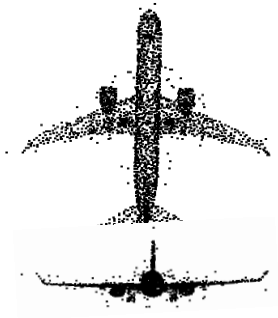
clean plane



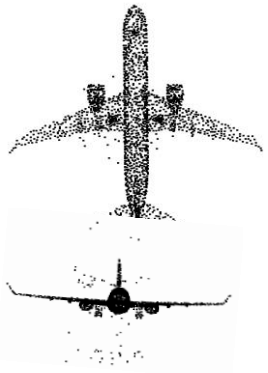
C&W L2 attack



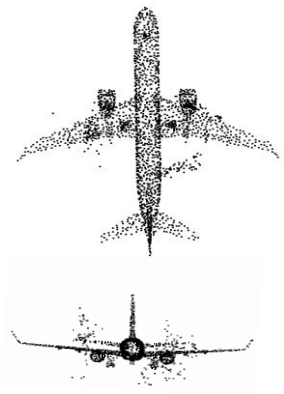
C&W chamfer attack



C&W hausdorff attack

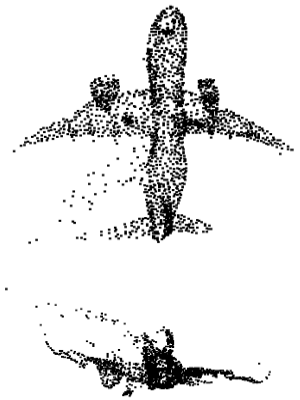


C&W cluster attack

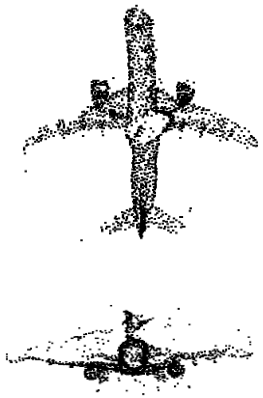


C&W object attack

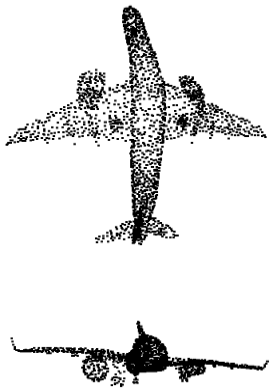
IFGM attack (to toilet)



Single-layered LG-GAN attack (to vase)



LG attack (to sofa)



LG-GAN attack (to lamp)

Results

	Target [4]	Defense (SRS) [39]	Defense (DUP-Net) [39]	ℓ_2 dist (meter)	Chamfer dist (meter)	Time (second)
C&W + ℓ_2 [33]	100	0	0	**0.01**	0.006	40.80
C&W + Hausdorff [33]	100	0	0	—	0.005	42.67
C&W + Chamfer [33]	100	0	0	—	**0.005**	43.73
C&W + 3 clusters [33]	94.7	2.7	0	—	0.120	52.00
C&W + 3 objects [33]	97.3	3.1	0	—	0.064	58.93
FGSM [17, 35]	12.2	5.2	2.8	0.15	0.129	0.082
IFGM [17, 35]	73.0	14.5	3.3	0.31	0.132	0.275
LG + Chamfer (ours)	96.1	75.4	13.9	0.63	0.137	0.037
single-layered LG-GAN (ours)	97.6	80.2	37.8	0.27	0.032	0.053
LG (ours)	97.1	85.0	72.0	0.25	0.028	**0.033**
LG-GAN (ours)	**98.3**	**88.8**	**84.8**	0.35	0.038	0.040

Table: Attack success rate (%, second to fourth column), distance (fifth-sixth column) between original sample and adversarial sample (meter per object) and generating time (second per object) on attacking PointNet. “Target” stands for white-box attacks. The hyper-parameter setting of two gray-box attacks is: for the simple random sampling (SRS) defense model, percentage of random dropped points is 60%~90%; for DUP-Net defense model, k = 50 and α = 0.9 from [39]. The default LG-GAN (ours) consists of multi-layered label embedding, ℓ_2 loss and GAN loss.

Thank You