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

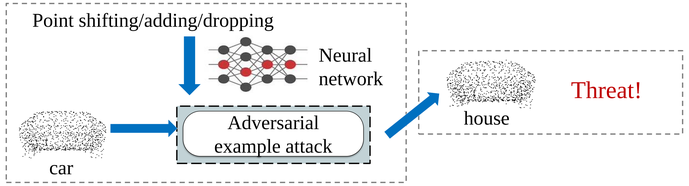
Hang Zhou[1*], Dongdong Chen[2*], Jing Liao[3], Kejiang Chen[1], Xiaoyi Dong[1], Kunlin Liu[1], Weiming Zhang[1], Gang Hua[4], Nenghai Yu[1]

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

IEEE/CVF 2020 Conference on Computer Vision and Pattern Recognition

CVPR SEATTLE WASHINGTON JUNE 16-18 2020

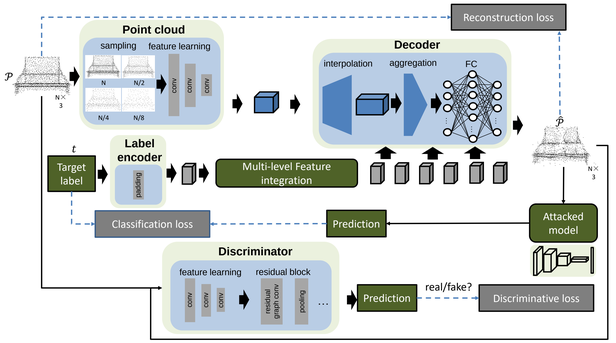
Introduction



Current attack methods:

- Optimization-based:
  Slow runtime/High attack success rate/Visible outliers
- Gradient-based:
  Fast runtime/Low attack success rate

Our Method



Encoder: $\overrightarrow{z_t} = (z_t^1, z_t^2, \dots, z_t^l) = \mathbf{E}_l(t)$

$\overrightarrow{F_{\mathcal{P}}} = (F_{\mathcal{P}}^1, F_{\mathcal{P}}^2, \dots, F_{\mathcal{P}}^l) = \mathbf{E}_{\mathcal{P}}(\mathcal{P})$

$\mathcal{P}_{adv} = \mathbf{D}_{\mathcal{P}}(\overrightarrow{z_t}, \overrightarrow{F_{\mathcal{P}}})$

Hierarchical Point Feature Learning: PointNet++ structure

Feature Decoding and Label Concatenation

$F_{\mathcal{P}}^{i+1''}(x) = \mathbf{FC}([z_t^i, F_{\mathcal{P}}^i]), \qquad i = 1,2,\dots,l-1$

Analysis

1. Residual graph convolutional block

graph convolution:

$$f_{out}(x) = w_0 f_{in}(x) + w_1 \sum_{q \in N(x)} f_{in}(q), \qquad \forall x \in v$$

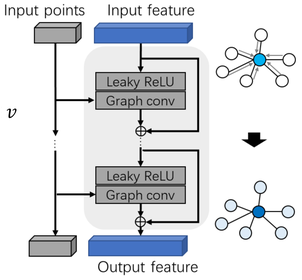


Fig. 1: The input points are utilized for querying the k nearest neighbors. The output feature of the block has the same dimension as the input feature.

2. Objective Loss Function

- loss of generator networks $\mathcal{G}$:
  $\mathcal{L}_{\mathcal{G}} = \mathcal{L}_{cls} + \alpha\mathcal{L}_{rec} + \beta\mathcal{L}_{dis}$
  where the classification loss: $\mathcal{L}_{cls} = -\left[t \log \mathcal{H}(\hat{\mathcal{P}}) + (1-t) \log\left(1 - \mathcal{H}(\hat{\mathcal{P}})\right)\right]$
  the reconstruction loss $\mathcal{L}_{rec}$ is the L2 norm
  and the discriminative loss: $\mathcal{L}_{dis}(\hat{\mathcal{P}}) = \|1 - \mathcal{D}_\theta(\hat{\mathcal{P}})\|_2^2$
- loss of discriminator networks $\mathcal{D}$:
  $\mathcal{L}_{\mathcal{D}}(\mathcal{P}, \hat{\mathcal{P}}) = \frac{1}{2}\|\mathcal{D}_\theta(\hat{\mathcal{P}})\|_2^2 + \frac{1}{2}\|1 - \mathcal{D}_\theta(\mathcal{P})\|_2^2$

3. Network structure

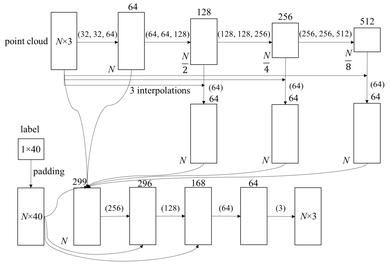


Fig. 2: The **generator part of the network architecture** of LG-GAN.

4. Difficult adversarial examples

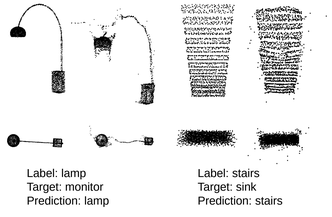


Fig. 3: Visualization of **difficult crafts with failed attacks**: objects from "lamp" and "stairs" categories and their adversarial examples.

Experiments

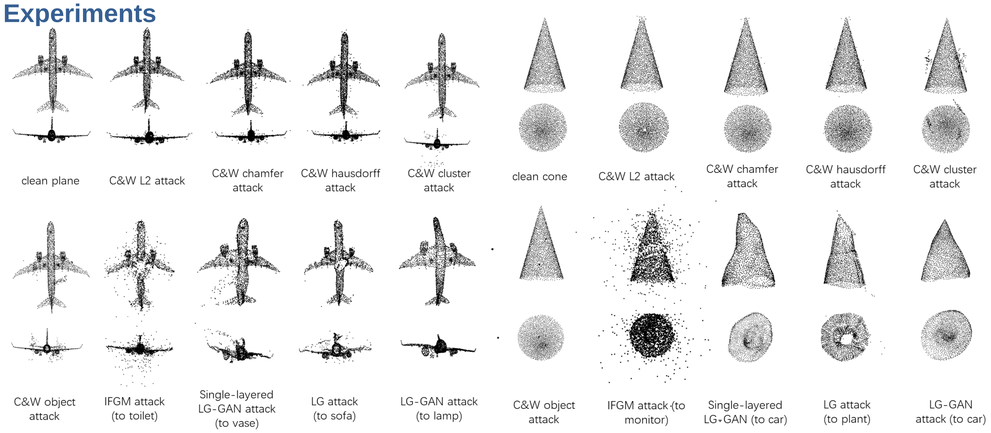


Fig. 4: **Qualitative results** on two point clouds from ModelNet40, "plane" and "cone".

	Target [5]	Defense (SRS) [43]	Defense (DUP-Net) [43]	ℓ_2 dist (meter)	Chamfer dist (meter)	Time (second)
C&W + ℓ_2 [36]	100	0	0	**0.01**	0.006	40.80
C&W + Hausdorff [36]	100	0	0	—	0.005	42.67
C&W + Chamfer [36]	100	0	0	—	**0.005**	43.73
C&W + 3 clusters [36]	94.7	2.7	0	—	0.120	52.00
C&W + 3 objects [36]	97.3	3.1	0	—	0.064	58.93
FGSM [20, 38]	12.2	5.2	2.8	0.15	0.129	0.082
IFGM [20, 38]	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 1: Attack success rate (%), distance, and generating time on **attacking PointNet from ModelNet40**. "Target" stands for white-box attacks.

ϵ (meter)	PointNet [3]	PointNet++ [4]	DGCNN [5]
0	88.6	89.5	87.9
0.01	88.5	89.5	87.8
0.1	77.2	89.6	87.2
0.5	13.2	89.7	57.9
1	3.5	86.4	14.4
2	1.7	61.9	4.1
10	2.0	6.0	2.6

Table 2: Detection accuracy (%) of point-cloud **translation attacks** on deep networks of ModelNet40. ϵ is the maximum stride size of translating one whole point cloud along X-axis, Y-axis and Z-axis.

Advantages

1. **Motivated by the limitations of existing 3D adversarial attack methods, we propose the first generation based adversarial attack method for 3D point-cloud recognition networks.**
2. **To support arbitrary-target attack, we design a novel label guided adversarial network "LG-GAN" by multiple intermediate feature incorporation.**
3. **Experiments on different recognition models demonstrate that our method is both more flexible and effective in targeted attack while being more efficient.**

Contact me:
Mail: zh2991@mail.ustc.edu.cn
Homepage:
http://home.ustc.edu.cn/~zh2991/

Looking for PostDoc position