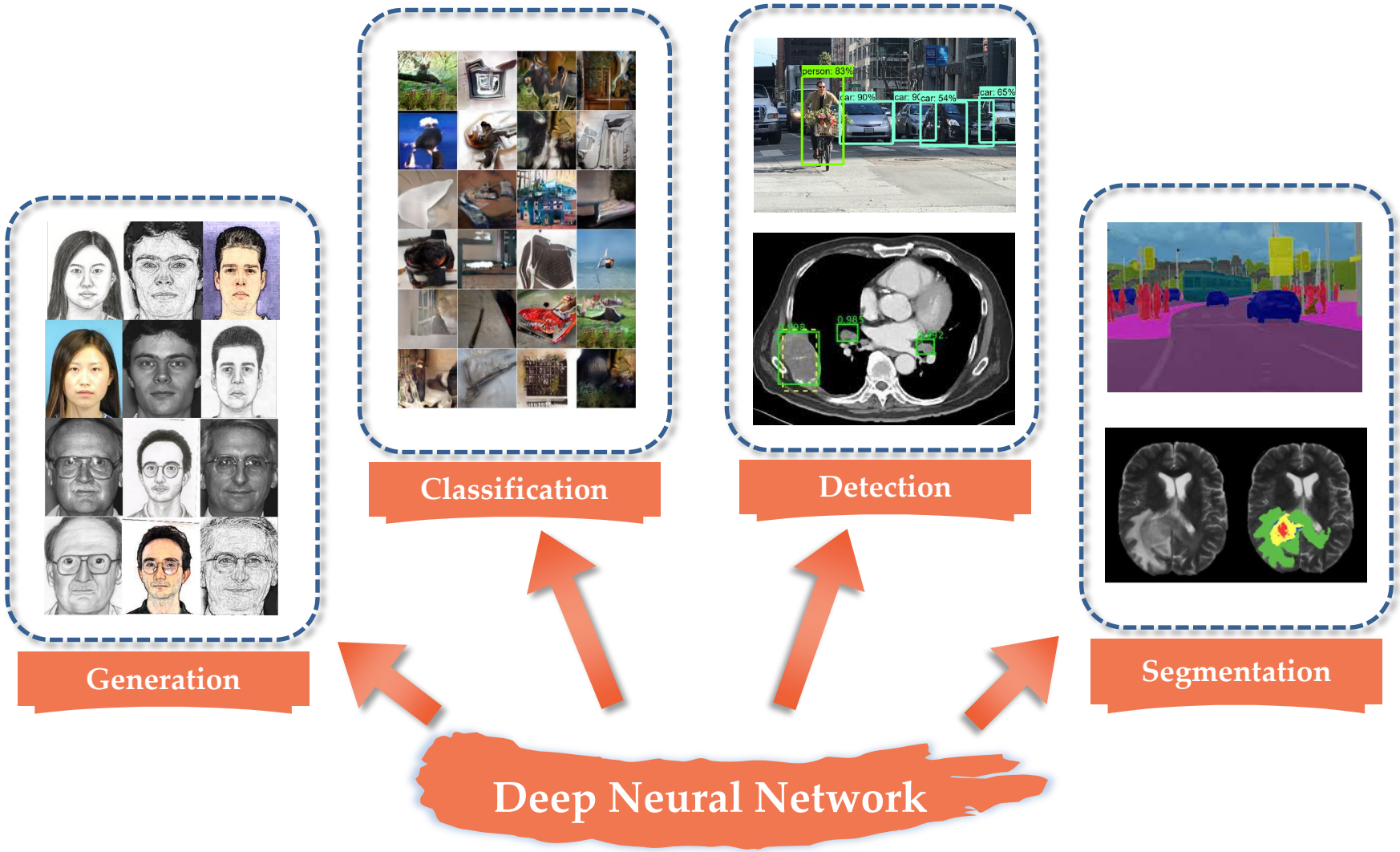
Modeling Adversarial Noise for Adversarial Training

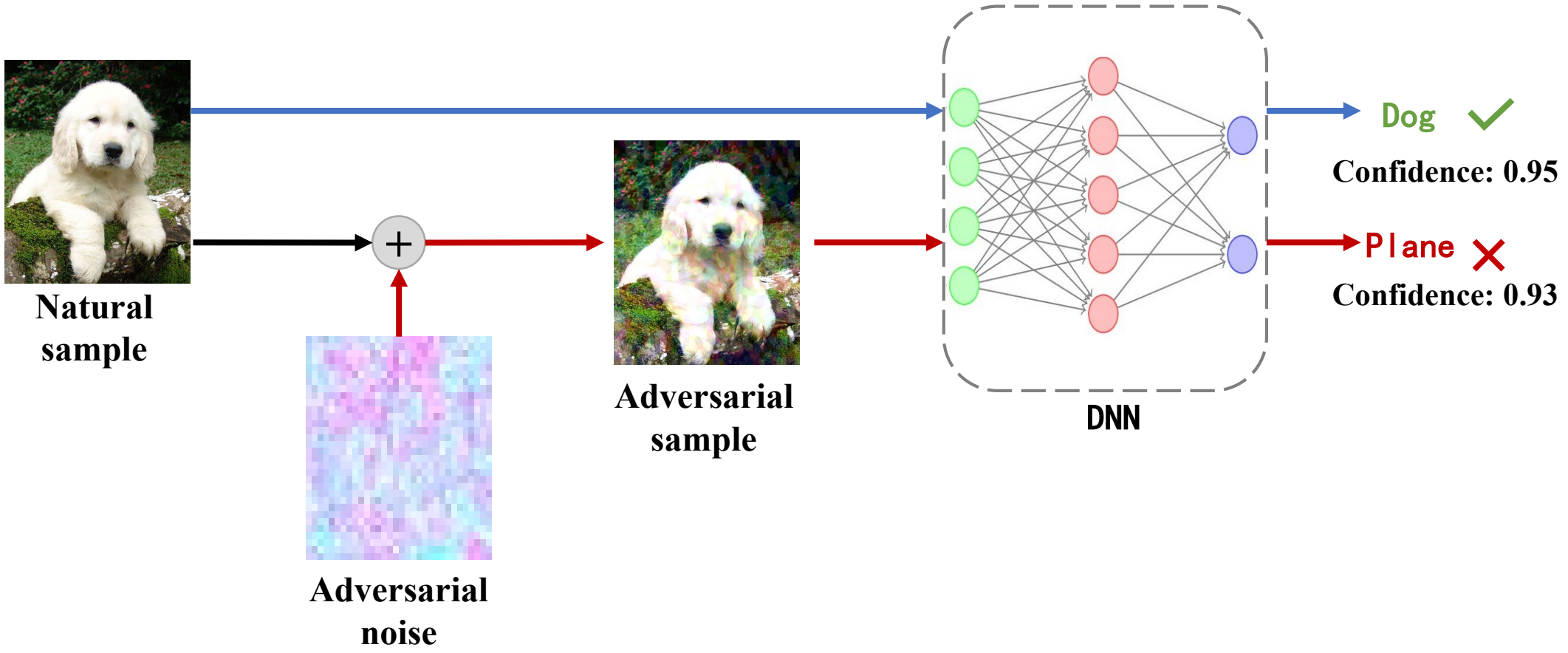
Dawei Zhou[1 2 *], Nannan Wang[1 *], Bo Han[3], Tongliang Liu[2 †]

[1] ISN Lab, Xidian University, [2] TML Lab, University of Sydney, [3]Hong Kong Baptist University

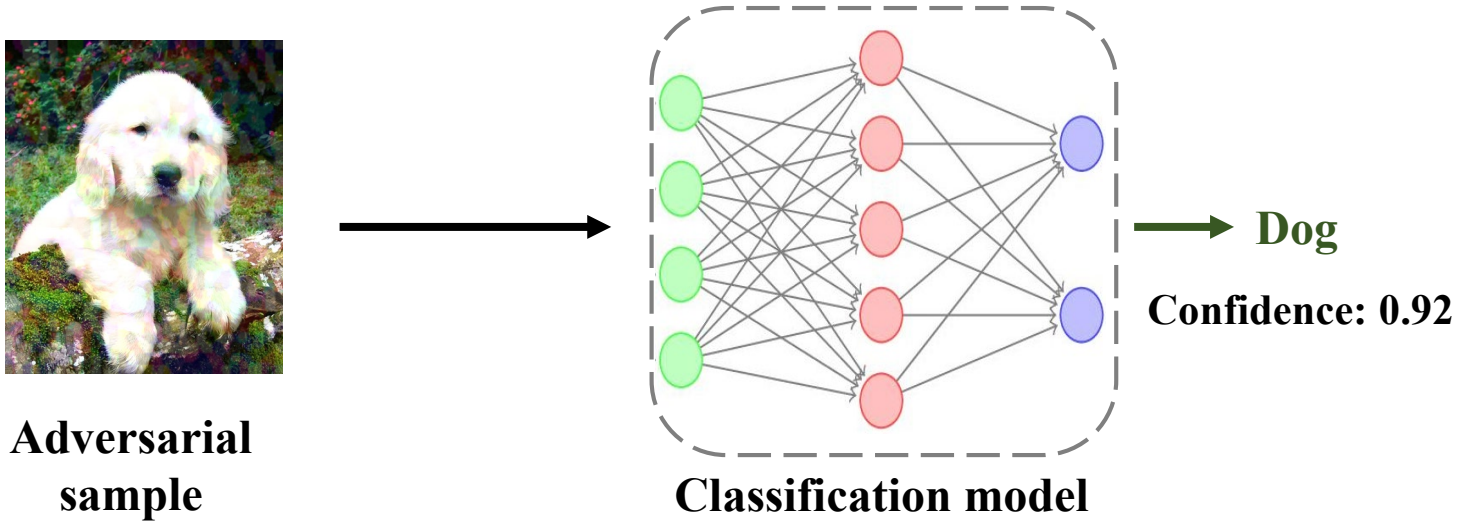
Deep neural networks have been widely used for multiple tasks



Deep neural networks are vulnerable to adversarial noise.



Introduction

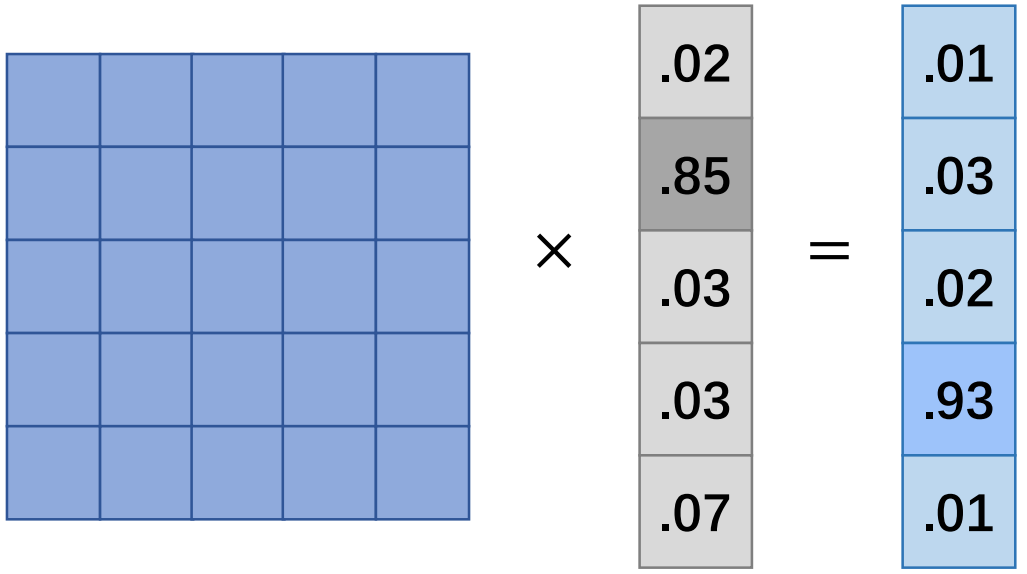


- Explicitly modeling the adversarial noise has not been considered.
- The relationship between the adversarial data and natural data has not been well studied yet.

Motivation

- Modeling the relationship between adversarial data and natural data can help infer natural data information by exploiting the adversarial data.
- Modeling adversarial noise in the low-dimensional label space can help avoid some high dimensionality problems.
- DNNs tend to use any available features including those that are well-generalizing, yet brittle. Adversarial noise can arise as a result of perturbing these features and it controls the flip from natural labels to adversarial labels.
- Adversarial noise thus contains well-generalizing features which can be modeled by learning the label transition from adversarial labels to natural labels.

Transition matirx



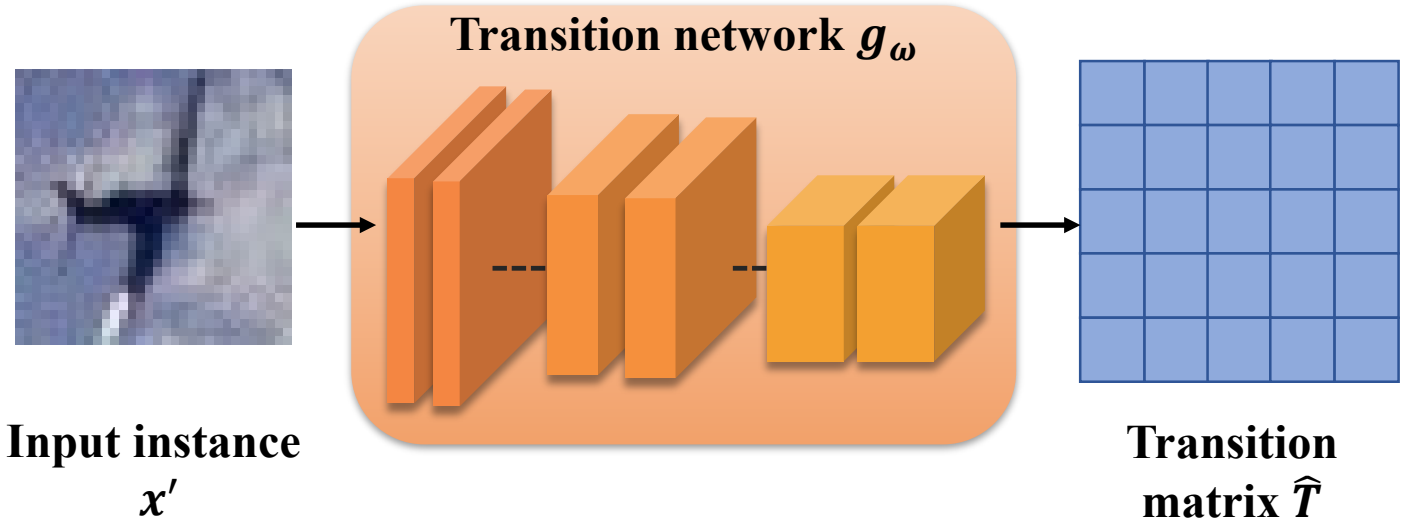
Transition matrix *T*

Adversarial Label *Y′*

Natural Label *Y*

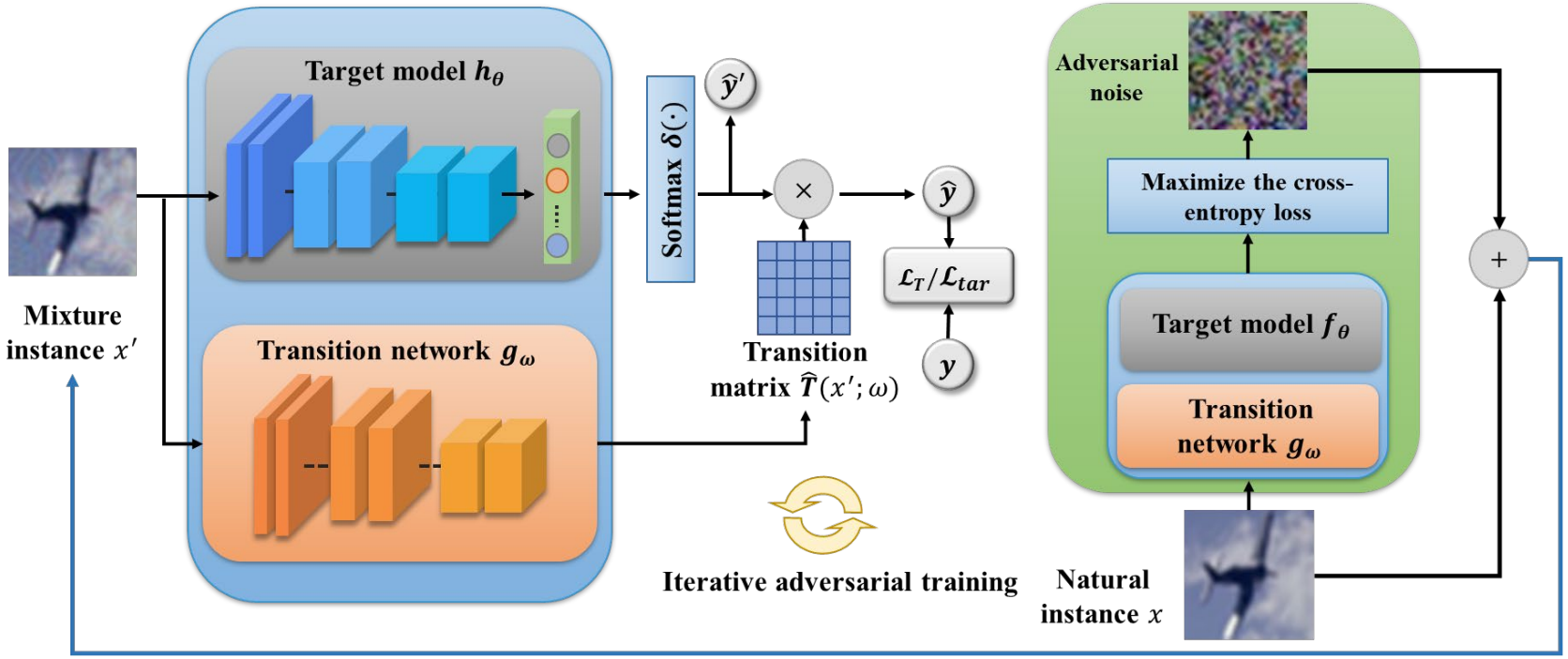
$$P(\boldsymbol{Y} \mid X' = x') = T(X' = x')^{\top} P(\boldsymbol{Y}' \mid X' = x')$$

Transition network



- We employ the transition network to estimate the label transition matrix.
- To optimize the parameter ω, we minimize the difference between the inferred natural labels and the ground-truth natural labels.

Training procedure



- Considering that adversarial data can be adaptively crafted, we conduct joint adversarial training on the target model and the transition network.

- Defending against adaptive attacks.

Table.1 Adversarial accuracy (higher is better) of MAN-based defense against adaptive attacks . The target model is ResNet-18. We report the results of the last checkpoint.

Dataset	Defense	None	PGD-40	AA	FWA-40
CIFAR-10	AT	**83.39**	42.38	39.01	15.44
	MAN	82.72	**44.83**	**39.43**	**29.53**
	TRADES	**80.70**	46.29	42.71	20.54
	MAN_TRADES	80.34	**48.65**	**44.40**	**29.13**
	MART	**78.21**	50.23	43.96	25.56
	MAN_MART	77.83	**50.95**	**44.42**	**31.23**
Tiny-ImageNet	AT	**48.40**	17.35	11.27	10.29
	MAN	48.29	**18.15**	**12.45**	**13.17**
	TRADES	**48.25**	19.17	12.63	10.67
	MAN_TRADES	48.19	**20.12**	**12.86**	**14.91**
	MART	**47.83**	20.90	15.57	12.95
	MAN_MART	47.79	**21.22**	**15.84**	**15.10**

- Defending against general attacks.

Table.2 Adversarial accuracy (higher is better) of MAN-based defense against general attacks on CIFAR-10. The target model is ResNet-18. We report the results of the last checkpoint.

Defense	None	PGD-40	AA	FWA-40	CW	DDN
MAN	89.01	81.07	79.90	80.02	77.89	77.82
Model_T	88.98	0.00	0.00	0.00	0.00	0.00

Future works

- We expect that our work can provide a new adversarial defense strategy for the community of adversarial learning.
- Optimize the MAN-based defense method to improve its transferability and its performance when applied to other adversarial training methods.
- How to effectively learn the transition matrix for the dataset with more classes is also our future focus.