LabelFool: A Trick In The Label Space

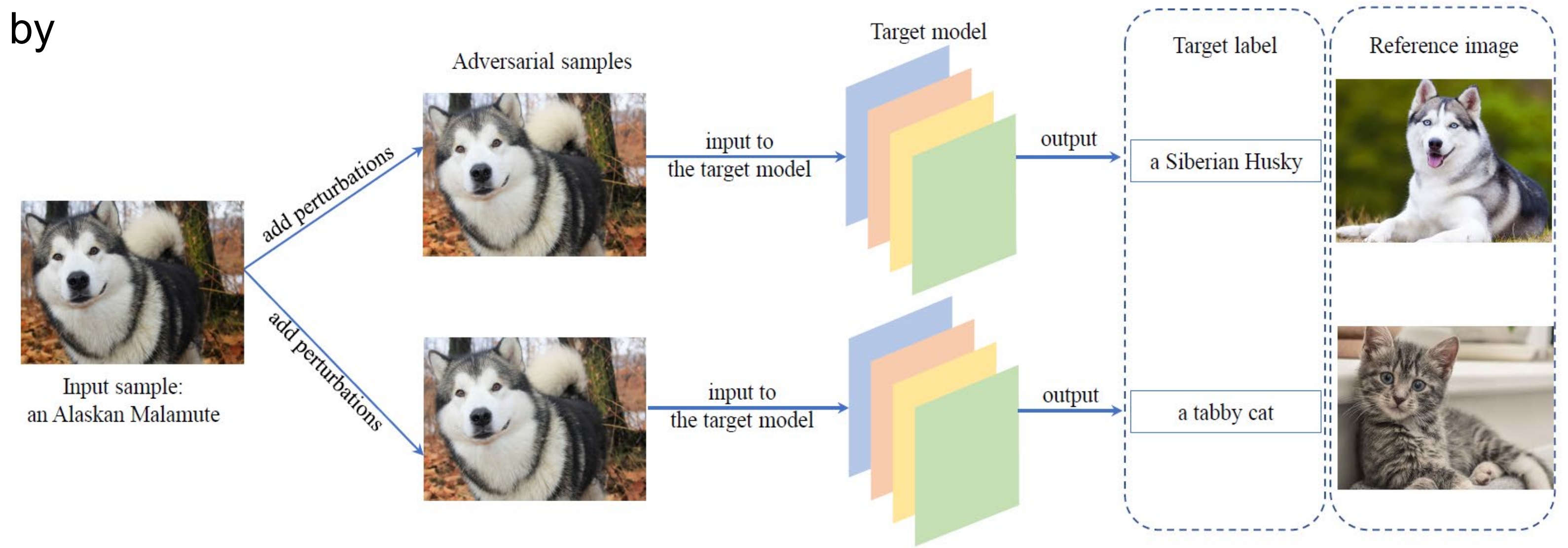
Yujia Liu (yujia_liu@pku.edu.cn), Tingting Jiang, Ming Jiang

IEEE WCCI 2022
IEEE WORLD CONGRESS ON COMPUTATIONAL INTELLIGENCE
PADUA, ITALY • 18 - 23 July, 2022

北京大學
PEKING UNIVERSITY

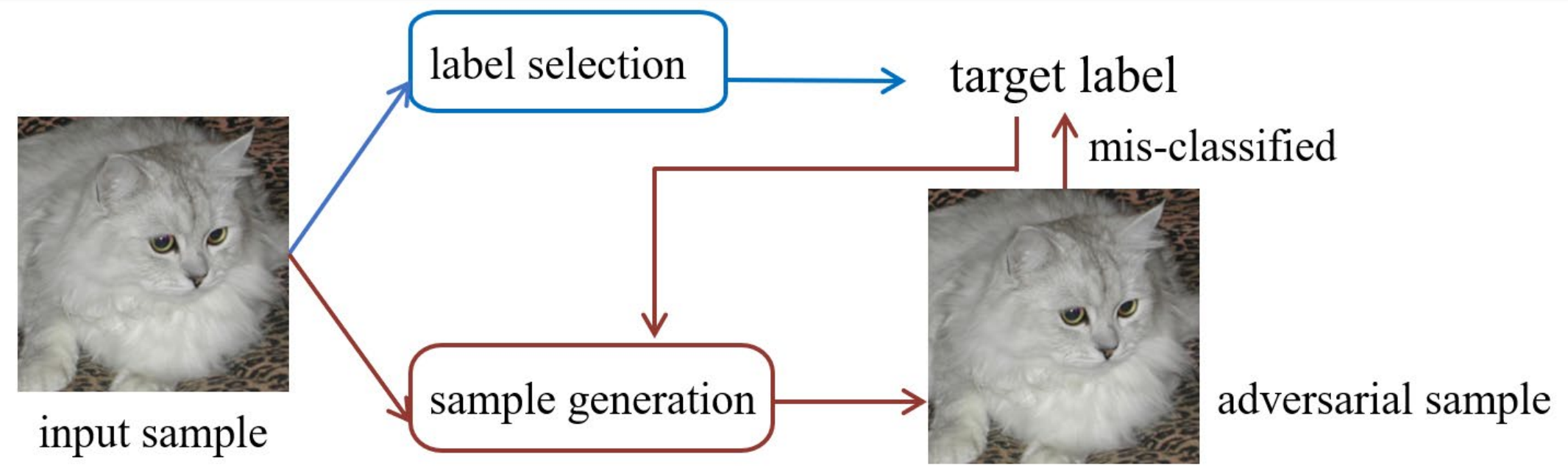
Motivation

- Imperceptibility of attacks in the label space is important in real applications with humans in loop, but it is overlooked by previous study.
- Annotations for target labels is time consuming so there needs an auto method.

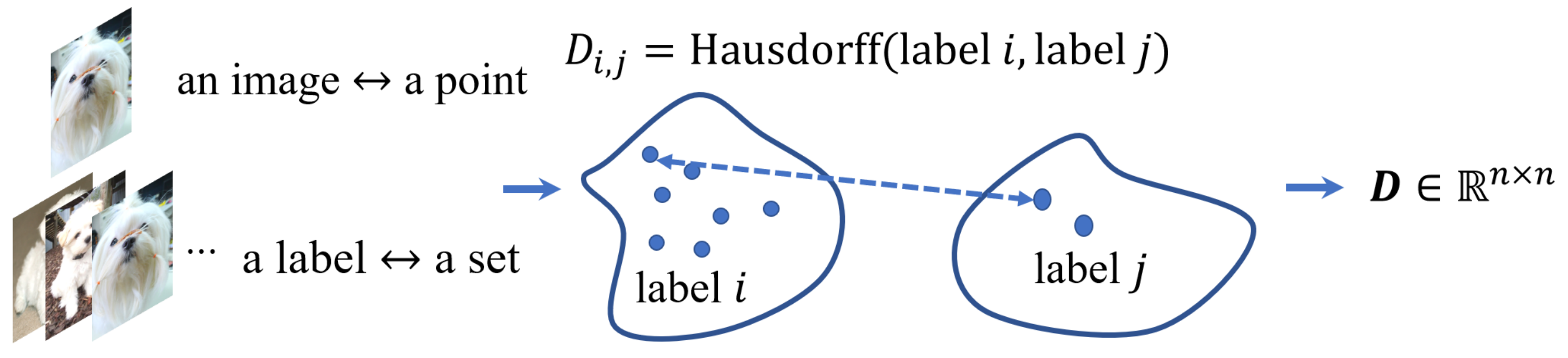


Method

LabelFool =
Label Selection + Sample Generation



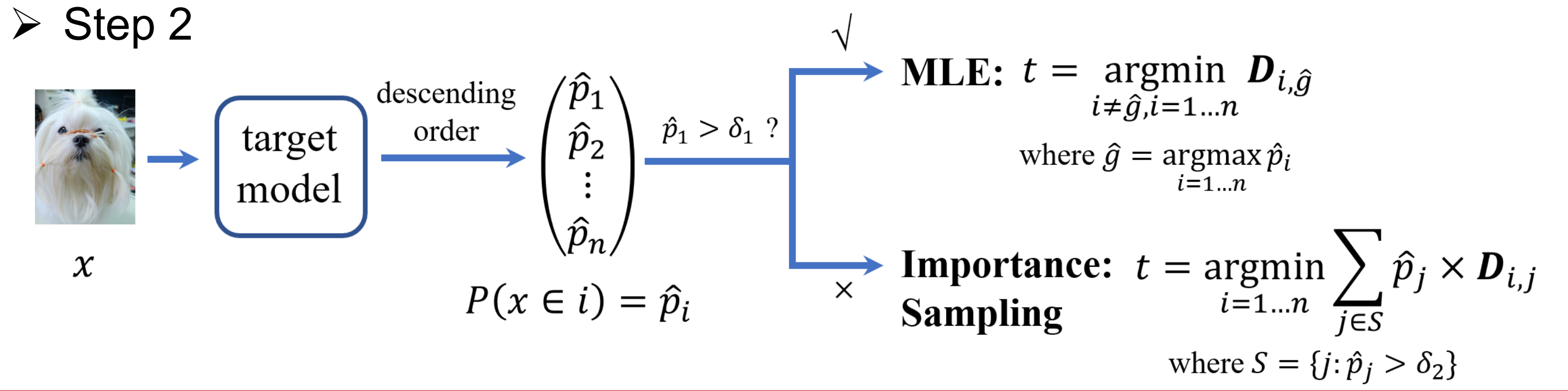
- Label Selection
  - Step 1



$$D_{i,j} = \text{Hausdorff}(\text{label } i, \text{label } j)$$

$$\boldsymbol{D} \in \mathbb{R}^{n \times n}$$

  - Step 2


x
target model
descending order
$\begin{pmatrix}\hat{p}_1\\ \hat{p}_2\\ \vdots\\ \hat{p}_n\end{pmatrix}$
$\hat{p}_1 > \delta_1$?
√
×

$$P(x \in i) = \hat{p}_i$$

MLE: $t = \underset{i \neq \hat{g}, i=1\ldots n}{\operatorname{argmin}} \boldsymbol{D}_{i,\hat{g}}$

where $\hat{g} = \underset{i=1\ldots n}{\operatorname{argmax}} \hat{p}_i$

Importance: Sampling $t = \underset{i=1\ldots n}{\operatorname{argmin}} \sum_{j \in S} \hat{p}_j \times \boldsymbol{D}_{i,j}$

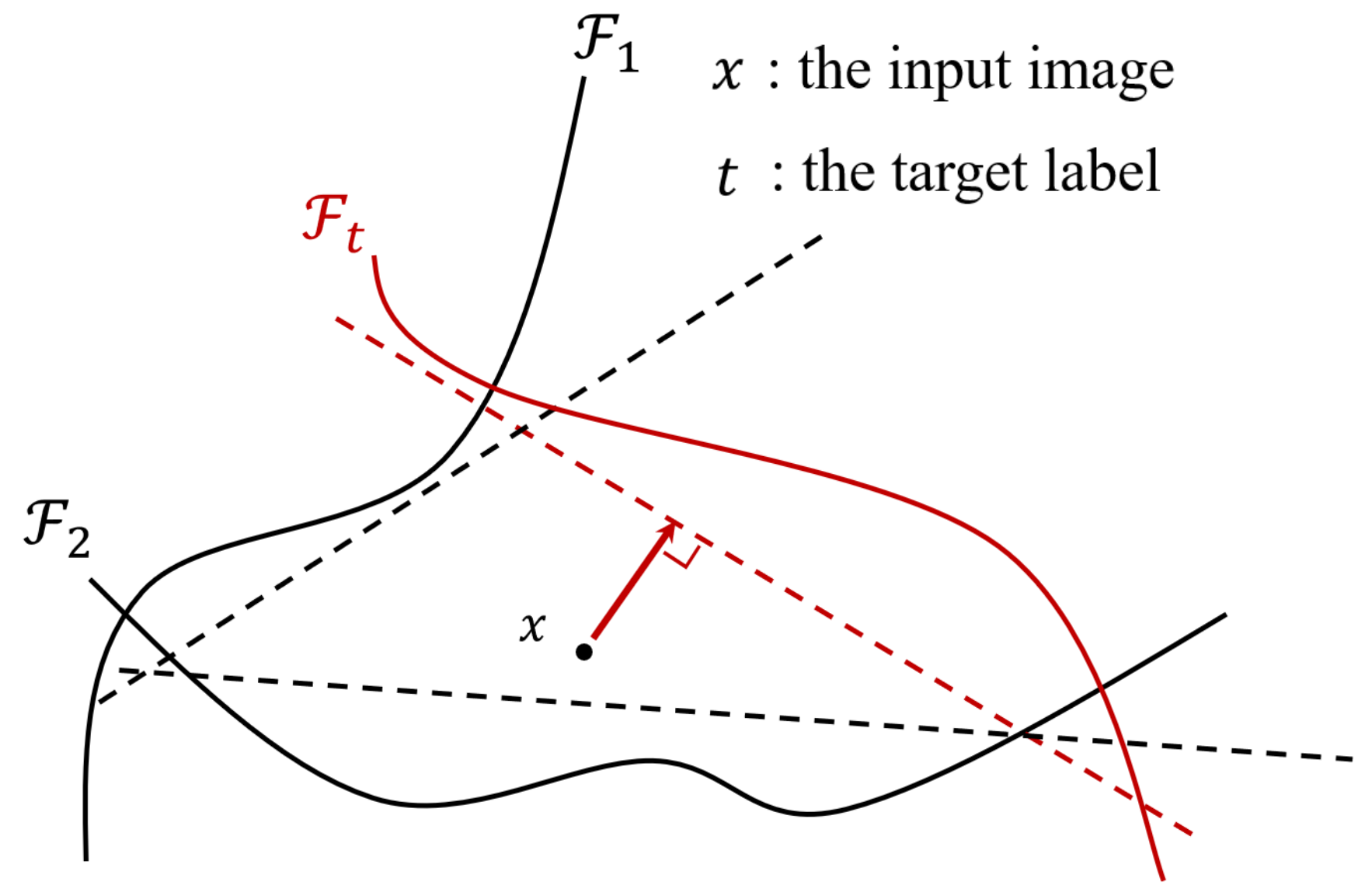
where $S = \{j : \hat{p}_j > \delta_2\}$

- Sample Generation

$\mathcal{F}_i$: the classification boundary between current class and class i

x : the input image

t : the target label



New Metrics

- Motivation: As no previous works have measured the imperceptibility of attacks in the label space, we propose new metrics based on subjective experiments to measure this.
- Confusion Rate (CR) measures the percentage of adversarial images whose target label successfully confuses a person.
- Real Confusion Rate (RCR) measures the percentage of adversarial images where the item corresponding to the target label does not appear, but the target label successfully confuses a person.

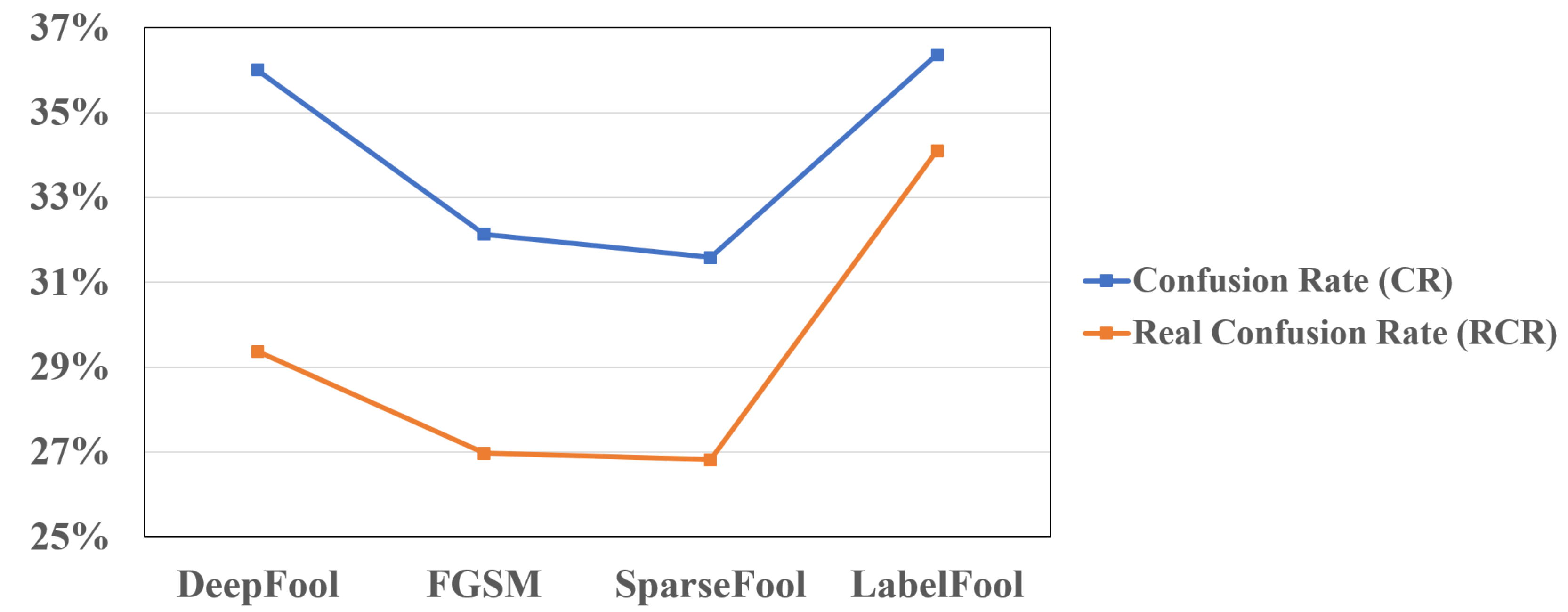
Experiments: Attack Rate

- Attack rate of different methods on different models.

Model	DeepFool	FGSM	SparseFool	LabelFool
Dataset: ImageNet				
ResNet-34	92.7%	95.0%	92.6%	**97.5%**
ResNet-50	93.1%	95.1%	92.5%	**97.9%**
VGG-19(bn)	92.0%	94.6%	83.7%	**97.5%**
AlexNet	90.4%	96.4%	89.1%	**97.4%**
Dataset: CASIA-WebFace				
SphereFace	98.7%	99.2%	97.8%	**99.3%**

Experiments: Imperceptibility in the Label & Image Space

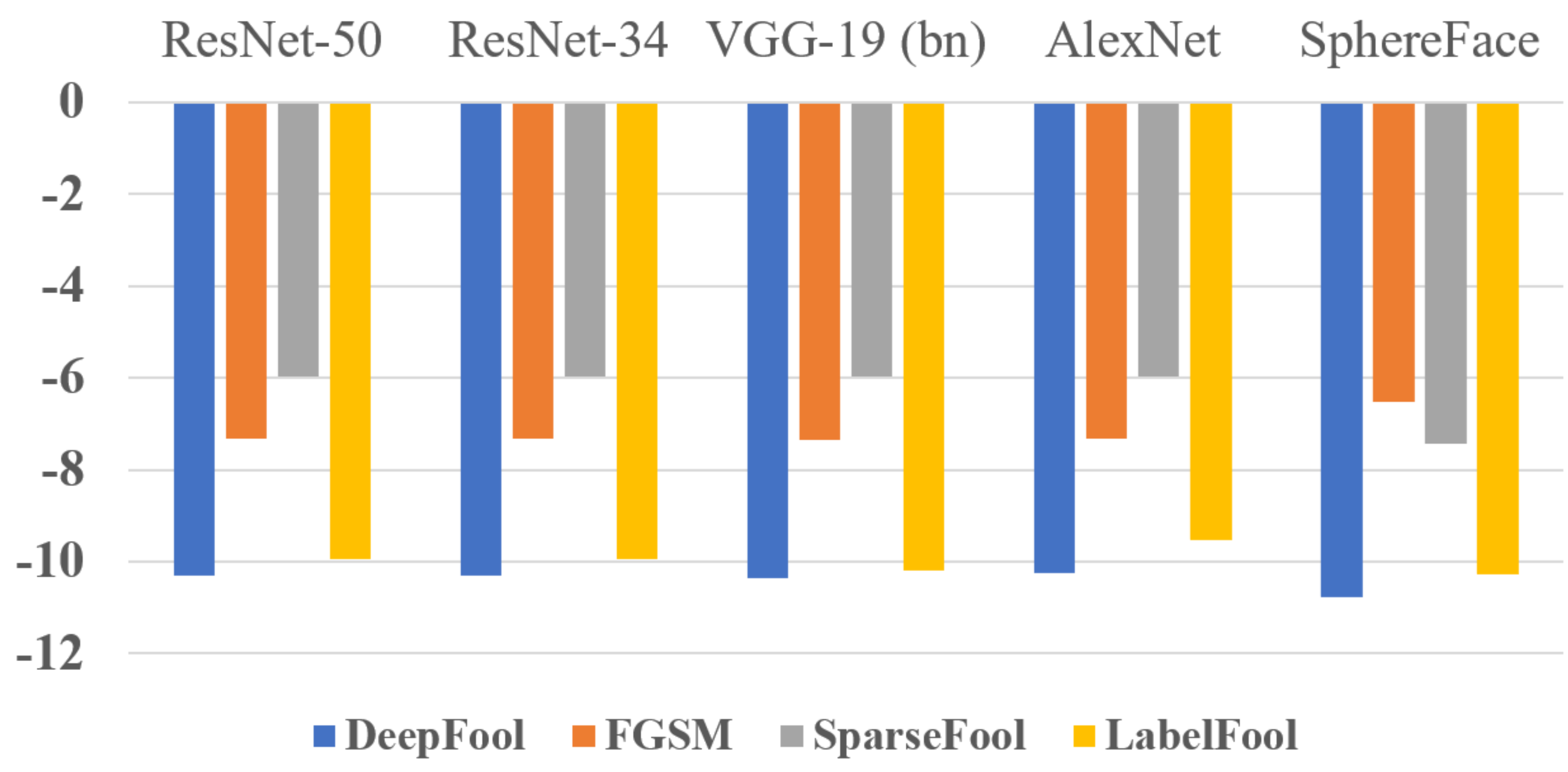
- Label Space: Average CR and RCR over 10 observers (3 females and 7 males, age between 20-29).



- Label Space: Visual results on CASIA-WebFace

Ground Truth ID & Input Sample	Target ID & Reference Image			
	DeepFool	FGSM	SparseFool	LabelFool
17	7319	6336	2609	1157

- Image Space: Log of the mean value of perceptibility for adversarial samples generated by different attack methods on different models.



Experiments: Different Label Selection Methods

- CR and RCR for different label selection models:

(1) Random: Select the target label uniformly at random.
(2) Easiest: Choose the second highest label of the output as the target label.
(3) Ours: Choose the target label by the proposed label selection method.

Method	Confusion Rate	Real Confusion Rate
Random	0.83%	-
Easiest	**49.17%**	38.00%
Ours	44.00%	**41.83%**

Conclusion

- Contribution:

(1) Observe the importance of imperceptibility of attacks in the label space and propose new metrics for it.
(2) Propose a feasible label selection method which achieves good imperceptibility in the label space.

- Limitation:

(1) Subjective experiments for CR and RCR is time consuming.
(2) Semantic distance is not considered.

Codes at
https://github.com/YogaLYJ/IJCNN2022_LabelFool.git !