BruSLeAttack

A Query-Efficient Score-based Black-box Sparse Attack

Quoc-Viet Vo, Ehsan Abbasnejad, Damith Ranasinghe

Presented by ***Quoc-Viet Vo***

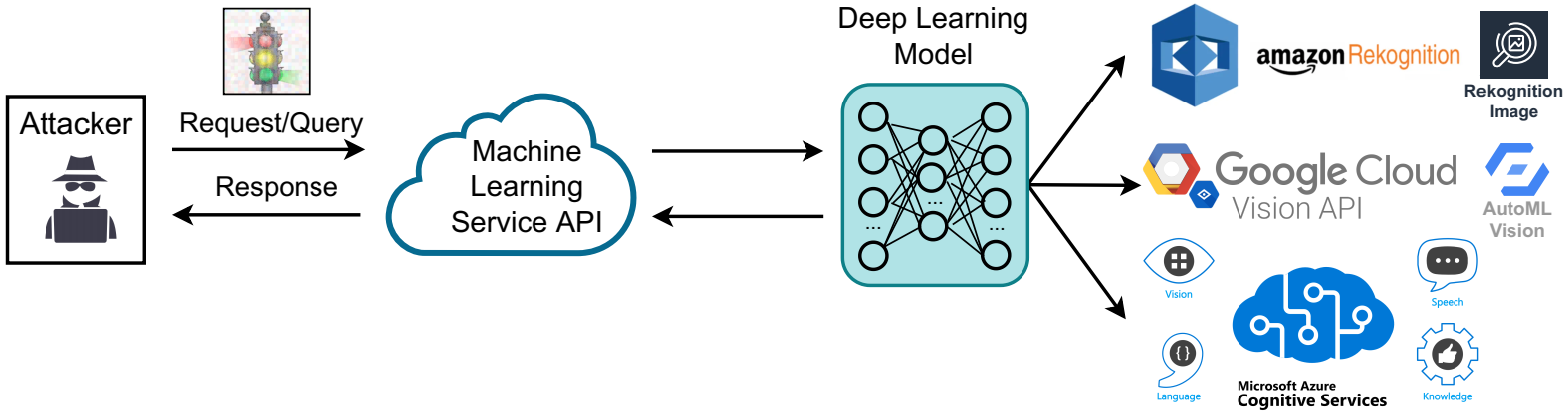
make history.

THE UNIVERSITY *of* ADELAIDE

Introduction

- **Dense Attacks** (L_2 or L_∞ norm): changing an entire image (widely explored).
- **Sparse Attacks** (L_0 norm): changing a few pixels (less well studied).

Motivation



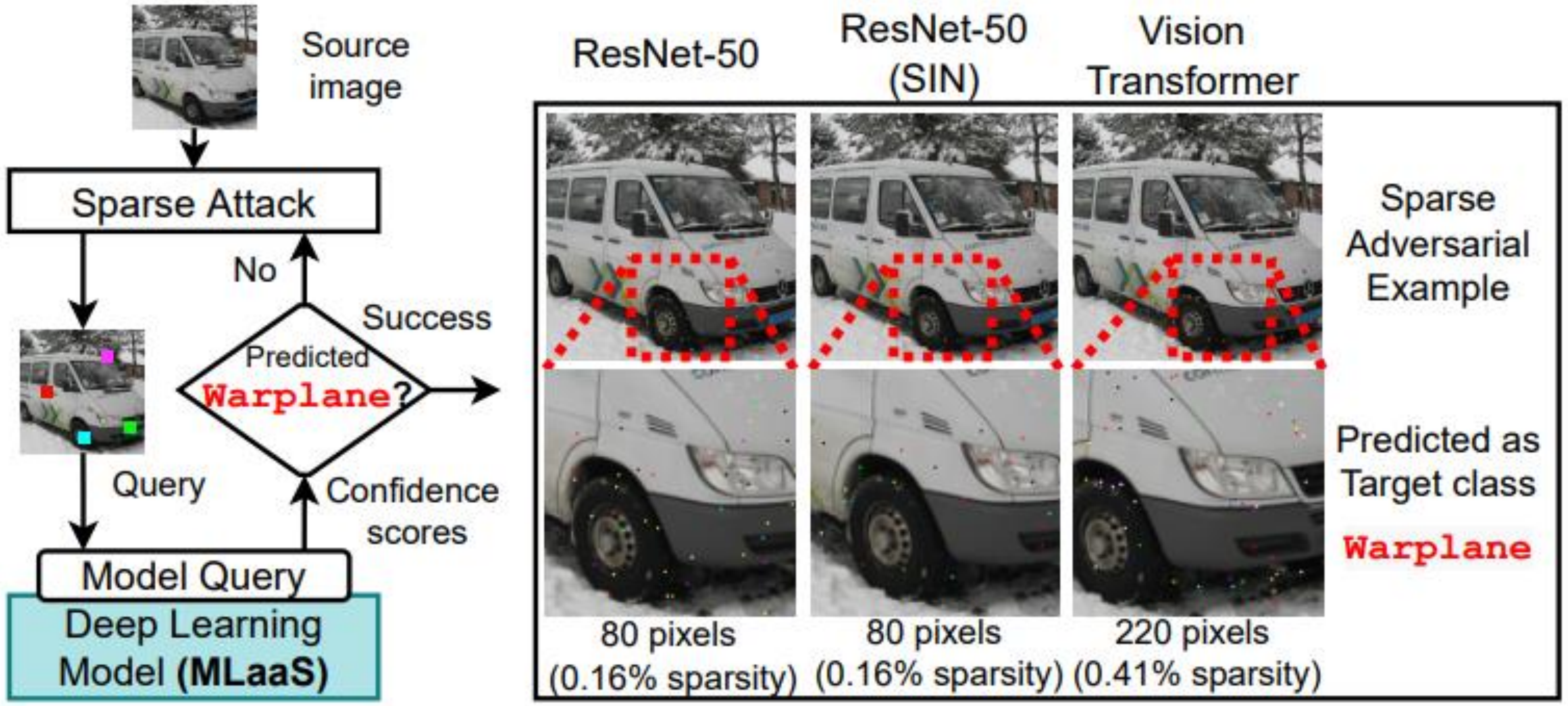
In real-world systems, the model is hidden from users except for the access to the model's response. Thus, it is a practical and threatening attack.

SCAN ME

THE UNIVERSITY of ADELAIDE

SCAN ME

BruSLeAttack against ImageNet



Threat model:

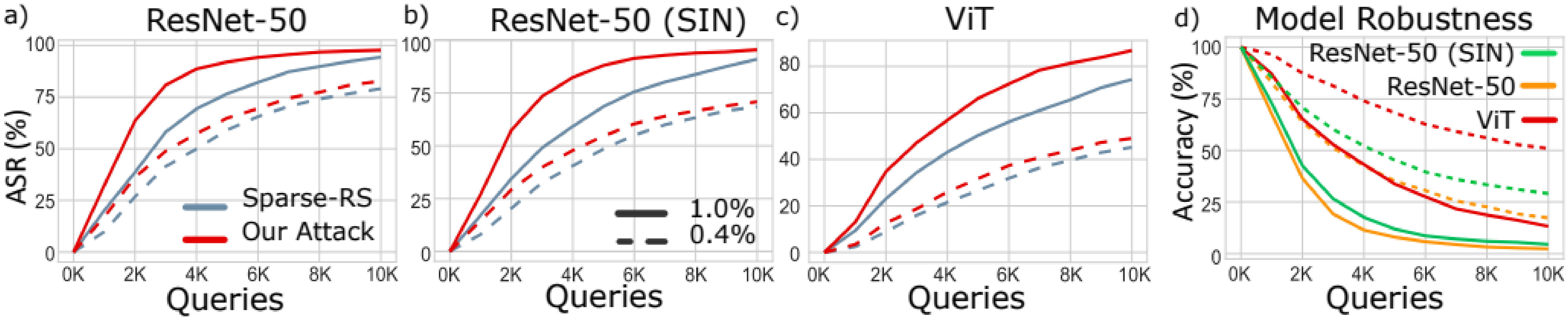
- Have access to output score (Score-based)
- Alter a few pixels (Sparse)

THE UNIVERSITY of ADELAIDE

BruSLeAttack against Deep Learning Models

SCAN ME

Attack Transformers & Convolutional Nets



Query Efficiency: within 10K queries, *BruSLeAttack* outperforms State-of-the-art *Sparse-RS* [4].

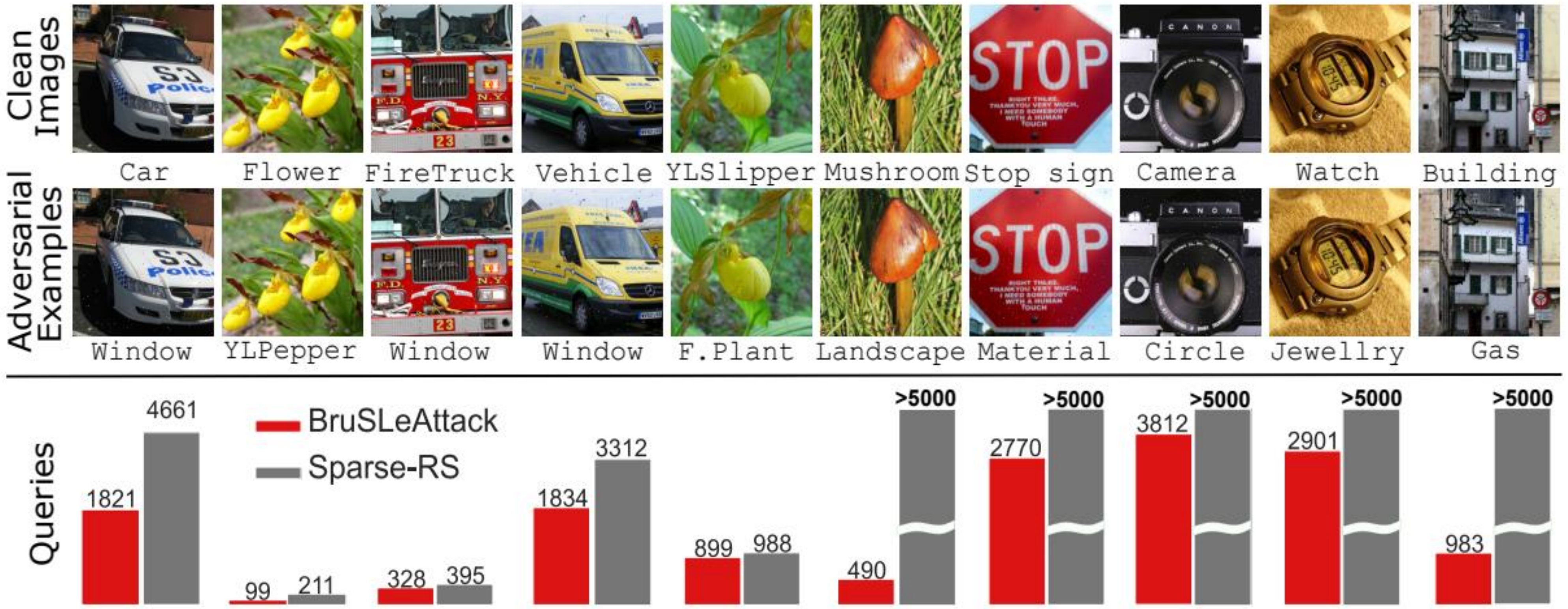
Attack Success Rate (ASR, up to 10K queries): *BruSLeAttack* achieves a much higher ASR than *Sparse-RS* across different query budgets.

[4] Francesco Croce, Maksym Andriushchenko, Naman D. Singh, Nicolas Flammarion, and Matthias Hein. Sparse-RS: A Versatile Framework for Query-Efficient Sparse Black-Box Adversarial Attacks. *Association for the Advancement of Artificial Intelligence (AAAI)*, 2022.

THE UNIVERSITY of ADELAIDE

BrusLeAttack against Google Cloud Vision API

Validation with Real-world system



Targeted Attack Against 10 Samples

BruSLeAttack is more query efficient than State-of-the-art Sparse-RS.

Project URL: https://brusliattack.github.io/

BruSLeAttack against Defended Models

SCAN ME

Sparsity	Undefended Model		l_∞-AT		l_2-AT		RND	
	SPARSE-RS	BRUSLEATTACK	SPARSE-RS	BRUSLEATTACK	SPARSE-RS	BRUSLEATTACK	SPARSE-RS	BRUSLEATTACK
0.04%	33.6%	**24.0%**	43.8%	**42.2%**	89.8%	**88.4%**	90.8%	**85.0%**
0.08%	13.2%	**6.8%**	26.8%	**24.4%**	81.2%	**79.2%**	82.2%	**72.6%**
0.12%	7.6%	**2.6%**	19.0%	**18.4%**	75.8%	**73.8%**	73.6%	**61.0%**
0.16%	5.2%	**1.0%**	16.6%	**14.8%**	71.4%	**69.2%**	64.8%	**51.4%**
0.2%	4.6%	**1.0%**	12.2%	**11.8%**	68.4%	**66.4%**	56.8%	**42.6%**

BruSLeAttack consistently outperforms *Sparse-RS* against different defense methods and different sparsity levels.

Challenges

- An NP-hard problem [1, 2].
- A discrete and non-differentiable search space (mixed discrete and continuous) [3].

[1] Modas and P. Moosavi-Dezfooli, S. Frossard. Sparsefool: a few pixels make a big difference. CVPR 2019.
[2] X. Dong, D. Chen, J. Bao, C. Qin, L. Yuan, W. Zhang, N. Yu, and D. Chen. GreedyFool: DistortionAware Sparse Adversarial Attack, NeurIPS, 2020.
[3] N. Carlini and D. Wagner. Towards evaluating the robustness of neural networks. IEEE SSP, 2017.

SCAN ME

Challenges

Problem formulation

$$\boldsymbol{x}^* = \arg\min_{\tilde{\boldsymbol{x}}} L(f(\tilde{\boldsymbol{x}}), y_{\text{target}}) \text{ s.t. } \|\boldsymbol{x} - \tilde{\boldsymbol{x}}\|_0 \leq B \,,$$

The search space is **width x height x channels x colors**.

- The search space is extremely enormous
- It is a significant barrier for attack progress

Achieving both query efficiency and a high attack success rate (ASR) for a high-resolution dataset is challenging.

THE UNIVERSITY of ADELAIDE

How hard is to discover sparse adversarial example in black-box settings?

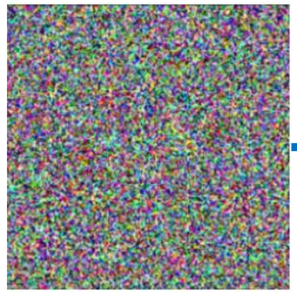
SCAN ME

Approach

An idea to reduce search space into **width x height**.

Source image

Synthetic color image



$$\tilde{\boldsymbol{x}} = \boldsymbol{u}\boldsymbol{x}' + (\mathbf{1} - \boldsymbol{u})\boldsymbol{x}$$

Image Maker

$$\boldsymbol{u} \in \{0,1\}^{w \times h}$$

Adversarial example

New formulation:

$$\boldsymbol{u}^* = \arg\min_{\boldsymbol{u}} \ell(\boldsymbol{u}) \quad \text{s.t. } \|\boldsymbol{u}\|_0 \leq B\,,$$

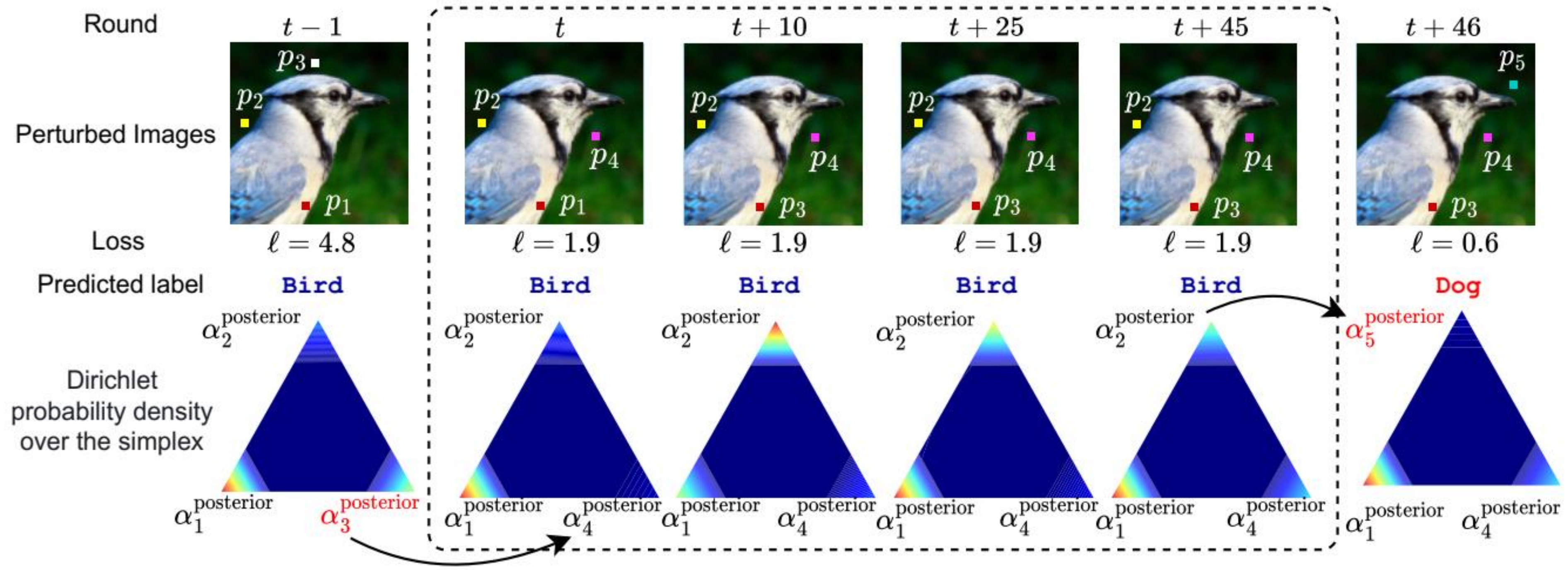
$$\ell(\boldsymbol{u}) := L(f(\boldsymbol{u}\boldsymbol{x}' + (\mathbf{1} - \boldsymbol{u})\boldsymbol{x}), y_{\text{target}})$$

THE UNIVERSITY of ADELAIDE

Approach

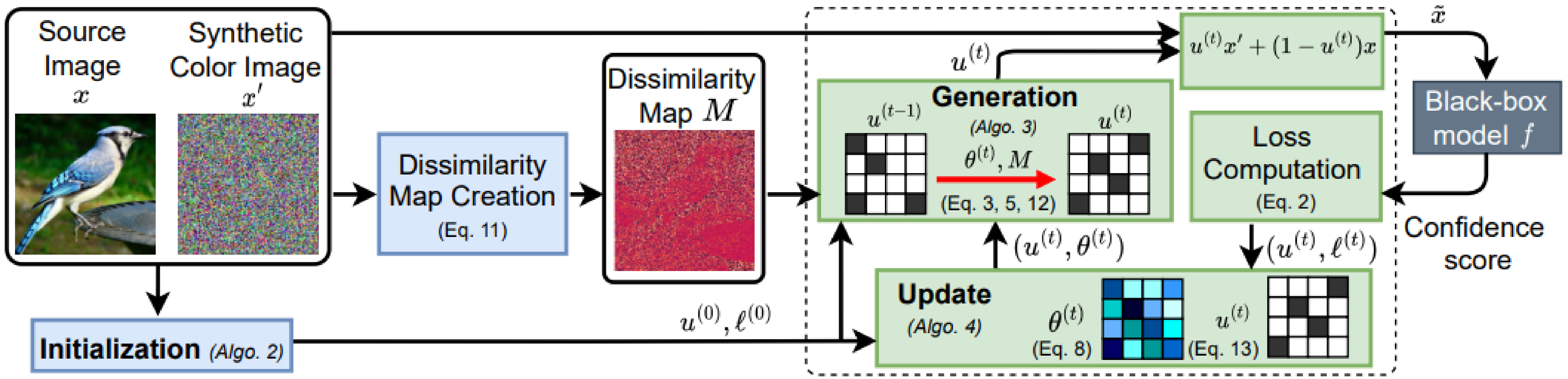
Employ **Bayesian Framework and history of pixel manipulation** to <u>learn the influence of pixels</u>.

Intuition: If a pixel has more impact on the model’s decision, replacing it is more likely to result in an increase in the loss. Thus, it should be less likely to be replaced.



Approach

BruSLeAttack algorithm



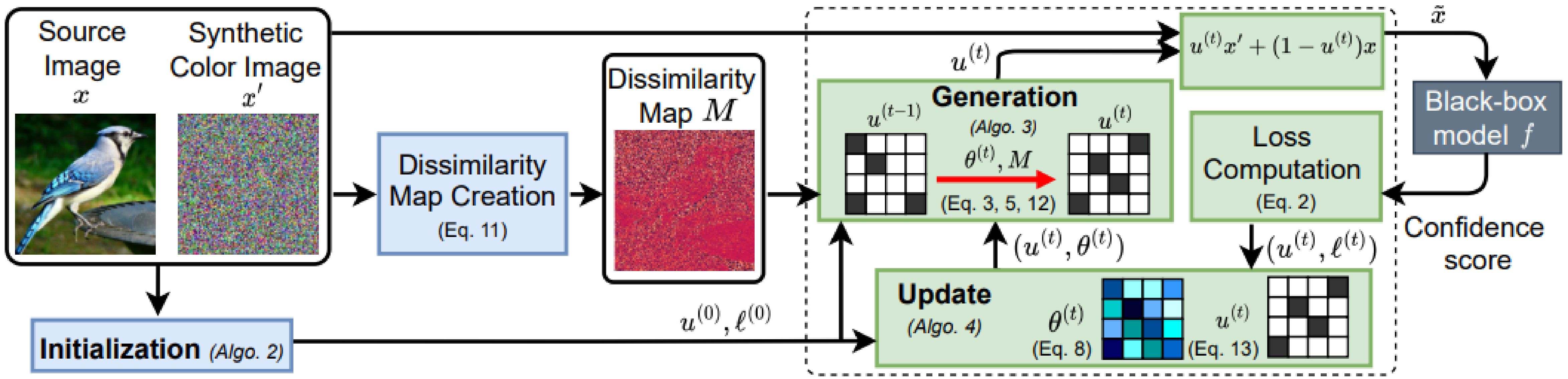
1. Create a dissimilarity map M
2. Initialize the some solutions randomly and choose the best $u^{(0)}$

SCAN ME

THE UNIVERSITY of ADELAIDE

Approach

BruSLeAttack algorithm



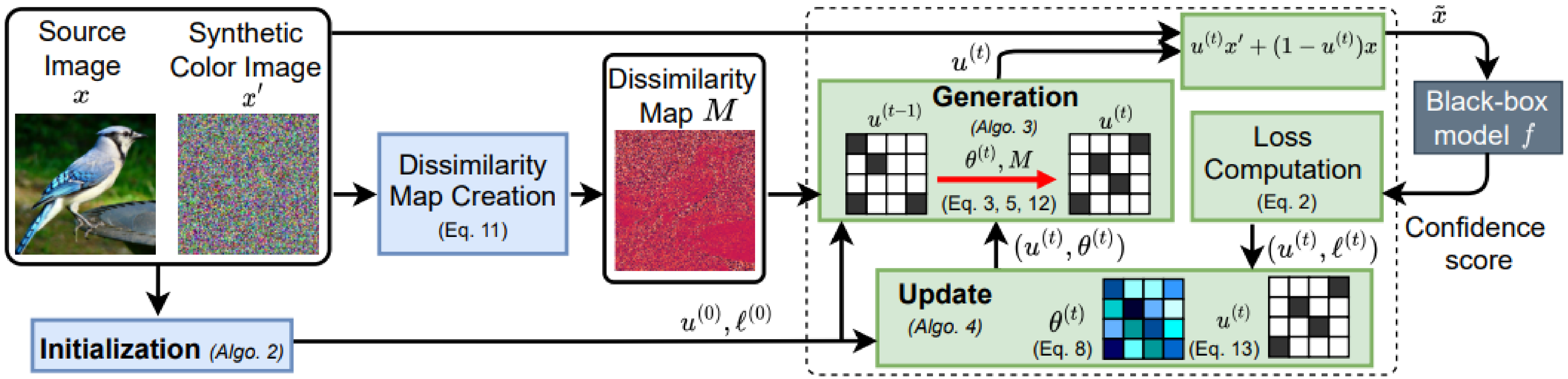
3. Sample new $u^{(t)}$ based on $\theta^{(t)}$ and M. Then craft an adversarial image $\tilde{x}$ from $u^{(t)}, x$ and x'.

4. Query model f and calculate loss $\ell^{(t)}$

Approach

SCAN ME

BruSLeAttack algorithm



5. Update both $\theta^{(t)}$ and $u^{(t)}$

Conclusion

BruSLeAttack

- Is capable of handling a discrete and non-differentiable search space.
- Is able to remedy the NP-hard problem.
- Is much more query-efficient than Sparse-RS in different benchmarks.

SCAN ME

THE UNIVERSITY of ADELAIDE

make history.

THE UNIVERSITY *of* ADELAIDE