INCHEON NATIONAL UNIVERSITY
COMPUTER VISION LABORATORY

IAP: Invisible Adversarial Patch Attack through Perceptibility-Aware Localization and Perturbation Optimization (ICCV 2025)

Undergraduate Researcher at CVLab
Lee Dohyeong

2025.11.14

Contents

- Introduction
- Related Work
- Method
- Experiments
- Conclusion

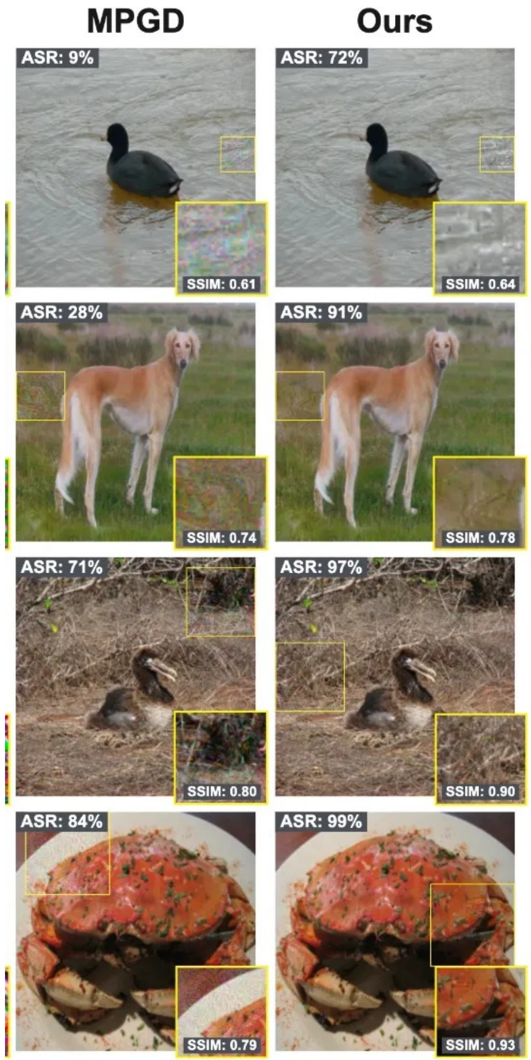
Introduction

- Traditional adv patches Problem :

  1. Fool the model
  2. Easily noticeable
  3. Low attack performance

- Contribution :

  - optimally balances the class localization and sensitivity scores
  - Restricting the changes in base color(reduces the saliency(두드러짐))
  - Demonstrates neutralize the latest patch defense techniques.



Related Work

1. Lack of Stealthiness (Too Salient)
   - Traditional patches (e.g., Google Patch) are visually overt and salient.
   - Result: Easily detected and blocked by recent defense mechanisms

2. The Trade-off (Invisibility vs. Efficacy)
   - Higher Invisibility → Lower Attack Efficacy.

→ Core Argument (IAP's Motivation)

- Constraining perturbation size (l_p - norm) makes targeted attacks impossible.

Method

- IAP pipeline

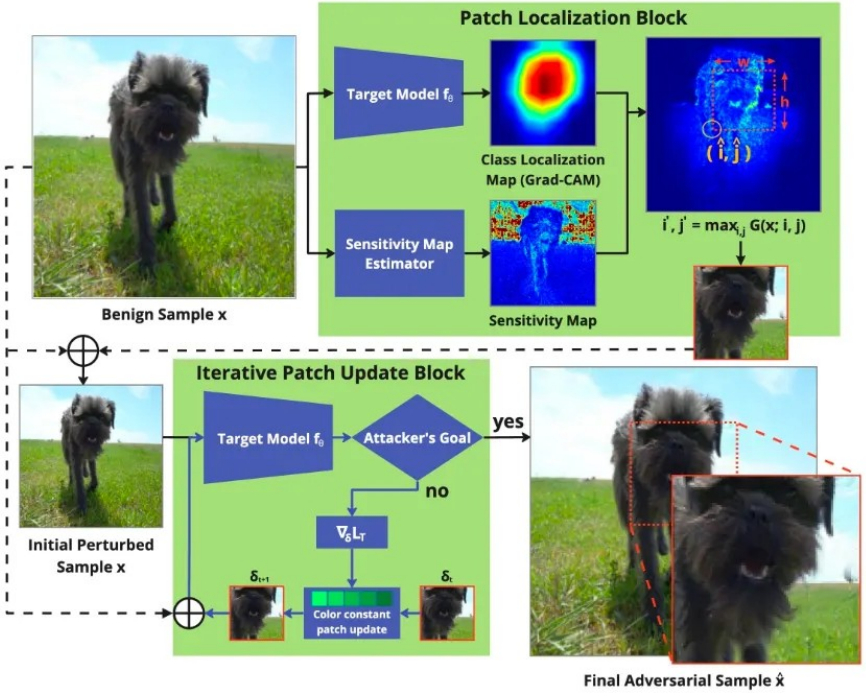


Figure 1. The overall pipeline of IAP for conducting targeted attacks with imperceptible adversarial patches, consisting of both patch localization and iterative patch update blocks.

- **Patch Localization Block**

  a. Target Model (AI) [Grad-CAM]

  Analyzes where the Target Model (AI) focuses for prediction.

  High score = High Attackability (Vulnerable region).

  b. Sensitivity Map Estimator

  Analyzes where Humans are insensitive to changes (complex textures).

  c. Combines (A) and (B) to find the location (i', j') that maximizes the trade-off.

- **Iterative Patch Update Block**

  a. Initialization : Start with the original image pixels (x)

  b. Query the Target Model: "Is the goal achieved?"

  c. Update using Averaged Gradients across RGB channels.

Method

- IAP

  1. Perception-aware placement

     : Targets model-vulnerable and human-insensitive regions

  2. Perturbation optimization

     : Uses perceptibility-aware loss and color-consistency update rules

- Attack Settings

  - White box(모델의 모든 조건을 아는 상태) + Targeted(A를 B로 정확히 속임)

Method

- **Location Selection Map**
  - $\hat{x}$ goal : $f_\theta(\hat{x}) = y_{targ}$

$$\hat{\boldsymbol{x}} = \underbrace{(1 - \boldsymbol{m}) \odot \boldsymbol{x}}_{\text{Background}} + \underbrace{\boldsymbol{m} \odot \boldsymbol{\delta}}_{\text{Sticker}}, \qquad (1)$$

$$\hat{\boldsymbol{x}} = \boldsymbol{x} +_{i,j} \boldsymbol{\delta}, \qquad (2)$$

x : *Original image*

$W * H * C$: *Width, Height, Color*

y : *Ground − Truth*

y_{targ}: *Target Attack label*

δ : *Adv Patch(pixel)*

i, j : *Patch starts location*

m : Mask[map]; 0, 1

Method

- Location Selection Map
  - Condition 1 : Vulnerable to AI
    - Focus condition 1 :
      - Advantages : High Attackability(easily fooled)
      - Disadvantage: High Human Sensitivity(Easily detected by human eyes.)
  - Condition 2 : Hide a lot
    - Focus condition 2 :
      - Advantage: High Invisibility(High Variance -> hide large perturbations)
      - Disadvantage: Low Attackability(attack "useless" region)

Method

- Location Selection Map
  - Goal : maximize G(optimal location score)

$$G(\boldsymbol{x};i,j)=\sum_{k=0}^{w}\sum_{l=0}^{h}\frac{J_y(\boldsymbol{x};i+k,j+l)}{\mathrm{Sens}(\boldsymbol{x};i+k,j+l)}, \quad (3)$$

f: *Victim Model*

$G(x;i,j)$: *Perturbation Priority Index*

$(i'j')$: *Optimal Position*

J_y: *AI Vulnerable map*

Sens: *human-sensitive map*

Method

- Location Selection Map
  - (4), (5) = Grad-CAM
    - Vulnerable to AI Map

$$\alpha_k^y = \frac{1}{u \times v} \sum_{i=0}^{u} \sum_{j=0}^{v} \frac{\partial g_\theta(\boldsymbol{x}, y)}{\partial A_{ij}^k}, \quad (4)$$

$$J_y(\boldsymbol{x}; i, j) = \mathrm{ReLU}\left(\sum_k \alpha_k^y \cdot A_{ij}^k \right). \quad (5)$$

A^k: k^{th} $feature\ piece$(Feature Map)

a_k^y: k^{th} $feature\ piece$(average pool)

g_θ: $AI\ decision\ score$

u,v : A^k Feature Map H, W

Method

Sensitivity Map

- Directional Standard Deviation
- Filtering Simple Edges ($\min$ operation)
- Inverse Relationship (Reciprocal)

$$\mathrm{Sens}(\boldsymbol{x}; i, j) = \frac{1}{\sigma_{ij} + \lambda}, \text{ where } \sigma_{ij} = \sqrt{\min(\sigma_{ij}^{x}, \sigma_{ij}^{y})}, \tag{6}$$

λ: *very small constant*

σ: *standard deviation*

Method

- Perturbation optimization

1. Regularized Adversarial Loss
2. Perceptibility Distance Metric (D)
   - Goal : min(D)

$$D(\boldsymbol{x}, \hat{\boldsymbol{x}}) = \frac{1}{h \times w} \sum_{k=i'}^{i'+w} \sum_{l=j'}^{j'+h} \mathrm{Sens}(\boldsymbol{x}; k, l) \cdot |x_{kl} - \hat{x}_{kl}|, \quad (7)$$

Method

- Loss Function

$$\mathcal{L}_T(\delta;\theta,x,y) = \underbrace{w_1 \cdot \mathcal{L}_{CE}(\hat{x}, y_{targ};\theta)}_{\text{Target Loss}} - \underbrace{w_2 \cdot \mathcal{L}_{CE}(\hat{x}, y;\theta)}_{\text{Original Loss}} + \underbrace{w_3 \cdot D(x,\hat{x})}_{\text{Recognition Penalty}}$$

Target Loss : $\{\hat{x}\ ,\ y_{targ}\}$ difference in probability

Original Loss : $\{\hat{x}\ ,\ y\}$ difference in probability

Recognition Penalty : visually detectable by human observers

Method

- Loss Function
  - Color Constant Update Rule (회색조 양자화(gray-level quantization))
    - Humans are indifferent changes in brightness if the Base Color remains the same

$$\boldsymbol{\delta}_{t+1} = \boldsymbol{\delta}_t - \eta \cdot \overline{\nabla_{\boldsymbol{\delta}}} \, \mathcal{L}_T(\boldsymbol{\delta}_t; \theta, \boldsymbol{x}, y) \odot \left(\boldsymbol{\delta}_t \oslash \mathrm{Sens}(\boldsymbol{x}) \right), \tag{9}$$

η: step size(learning rate)

$\overline{\nabla}_{\delta}\mathcal{L}_{\mathcal{T}}$: average Graident

Method

- Loss Function

Algorithm 1 Invisible Adversarial Patches (IAP)

Input: benign example $(\boldsymbol{x}, y)$, target class y_{targ}, victim model f_θ, and parameters s, T, η, w, h
    $J_y(\boldsymbol{x}) \leftarrow$ compute the class localization map of $\boldsymbol{x}$ based on Equation 5
    $\mathrm{Sens}(\boldsymbol{x}) \leftarrow$ compute the sensitivity map of $\boldsymbol{x}$ based on Equation 6
    $(i', j') \leftarrow$ find the optimal patch location based on Equation 3
    $\boldsymbol{m} \leftarrow$ define the mask indexed by (i', j') with patch size $w \times h$
    Initialize $\boldsymbol{\delta}_0 \leftarrow \boldsymbol{x}$
    for $t = 0, 1, \ldots T - 1$ **do**
        if prediction confidence $f_\theta(y_{\text{targ}} | \hat{\boldsymbol{x}}) \geq s$ **then**
            return $\hat{\boldsymbol{x}}$
        else
            $\mathcal{L}_T \leftarrow$ define the total adversarial loss function based on Equation 8
            $\boldsymbol{\delta}_{t+1} \leftarrow \boldsymbol{\delta}_t - \eta \cdot \overline{\nabla_{\boldsymbol{\delta}}}\, \mathcal{L}_T(\boldsymbol{\delta}_t; \theta, \boldsymbol{x}, y) \odot \left(\boldsymbol{\delta}_t \oslash \mathrm{Sens}(\boldsymbol{x})\right)$
            $\boldsymbol{\delta}_{t+1} \leftarrow \mathrm{clip}(\boldsymbol{\delta}_{t+1}, 0, 1)$
        $\hat{\boldsymbol{x}} \leftarrow \boldsymbol{x} +_{i', j'} \boldsymbol{\delta}_{t+1}$
Output: $\hat{\boldsymbol{x}}$

Experiments

- Comparison with SOTA Methods

Dataset	Method	ResNet-50			VGG 16			Swin Transformer Tiny			Swin Transformer Base		
		ASR	LPIPS$_L$(↓)	SSIM$_L$(↑)	ASR	LPIPS$_L$(↓)	SSIM$_L$(↑)	ASR	LPIPS$_L$(↓)	SSIM$_L$(↑)	ASR	LPIPS$_L$(↓)	SSIM$_L$(↑)
ImageNet	Google Patch	99.10	0.74	0.010	100.0	0.76	0.002	99.80	0.77	0.002	97.9	0.77	0.003
	LaVAN	100.0	0.78	0.010	93.60	0.79	0.002	99.70	0.78	0.005	100.0	0.78	0.004
	GDPA	93.70	0.57	0.350	89.20	0.61	0.310	83.70	0.54	0.390	85.10	0.54	0.360
	MPGD	97.80	0.24	0.790	96.50	0.32	0.810	98.80	0.19	0.800	70.50	0.20	0.800
	IAP	99.50	0.12	0.940	99.10	0.23	0.900	99.60	0.06	0.980	99.40	0.07	0.970
VGG Face	Google Patch	97.73 ± 1.56	0.75 ± 0.03	0.01 ± 0.00	99.97 ± 0.05	0.87 ± 0.01	0.00 ± 0.00	99.13 ± 0.17	0.81 ± 0.02	0.01 ± 0.02	97.90 ± 0.50	0.89 ± 0.04	0.00 ± 0.00
	LaVAN	99.00 ± 1.41	0.81 ± 0.04	0.01 ± 0.00	99.83 ± 0.24	0.86 ± 0.01	0.00 ± 0.00	100.0 ± 0.00	0.85 ± 0.00	0.01 ± 0.00	99.70 ± 0.29	0.85 ± 0.00	0.00 ± 0.00
	GDPA	99.07 ± 0.76	0.62 ± 0.03	0.31 ± 0.02	95.71 ± 3.28	0.62 ± 0.07	0.31 ± 0.12	95.10 ± 3.47	0.61 ± 0.03	0.33 ± 0.01	72.41 ± 12.6	0.63 ± 0.06	0.29 ± 0.10
	MPGD	67.11 ± 10.8	0.38 ± 0.00	0.61 ± 0.00	86.90 ± 1.61	0.42 ± 0.02	0.65 ± 0.02	95.52 ± 0.54	0.37 ± 0.01	0.64 ± 0.00	91.20 ± 7.45	0.38 ± 0.01	0.61 ± 0.01
	IAP	94.53 ± 3.06	0.21 ± 0.03	0.90 ± 0.01	99.44 ± 0.49	0.21 ± 0.01	0.92 ± 0.01	99.07 ± 0.33	0.26 ± 0.01	0.86 ± 0.00	98.20 ± 0.86	0.28 ± 0.02	0.86 ± 0.02

Table 1. Comparisons of ASR (%) and imperceptibility between different adversarial patch attacks on VGG Face. For MPGD, we consider perturbations bounded by $\epsilon = 16/255$ in ℓ_∞-norm. The subscripts $_L$ and $_G$ represent the imperceptibility measures at local and global scales, respectively. Note that a lower LPIPS score indicates the generated adversarial patches are less perceptible.

ASR : Attack Success Rate (공격 성공률)

LPIPS : Imperceptibility (화질 저하 점수)

SSIM : Similar two images are in luminance, contrast, and structure.(1=유사)

Experiments

- Visual Quality Assessment

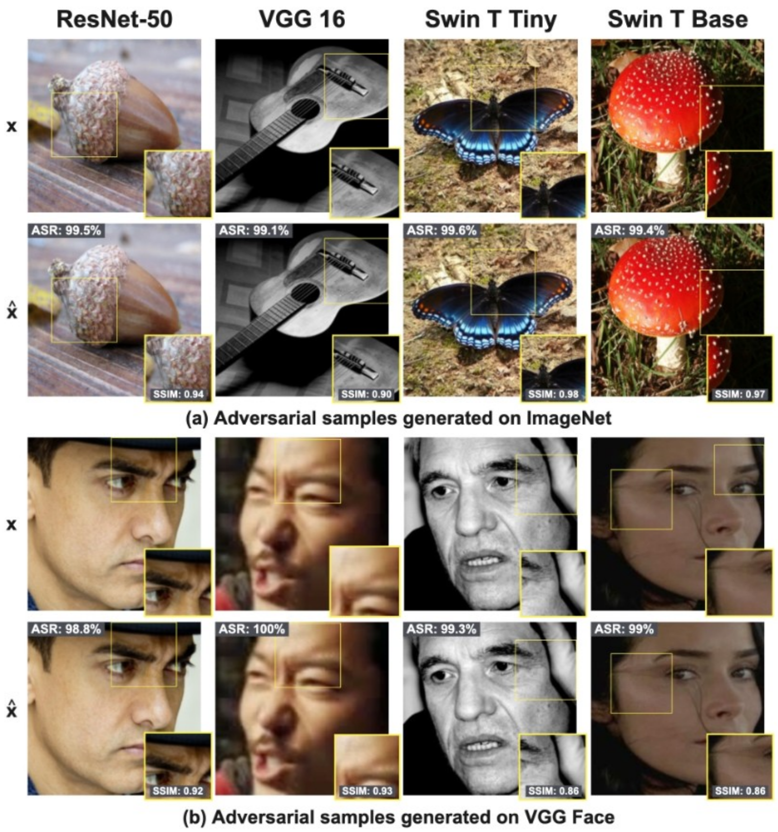


Figure 2. Visualizations of original images (x) and their adversarial counterparts ($\hat{x}$) generated by IAP. The smaller images in the bottom-right corner indicate the optimal location (i', j').

Experiments

- Human Perceptibility Study
  - Participants: 28 ML experts (Familiar with adversarial attacks)

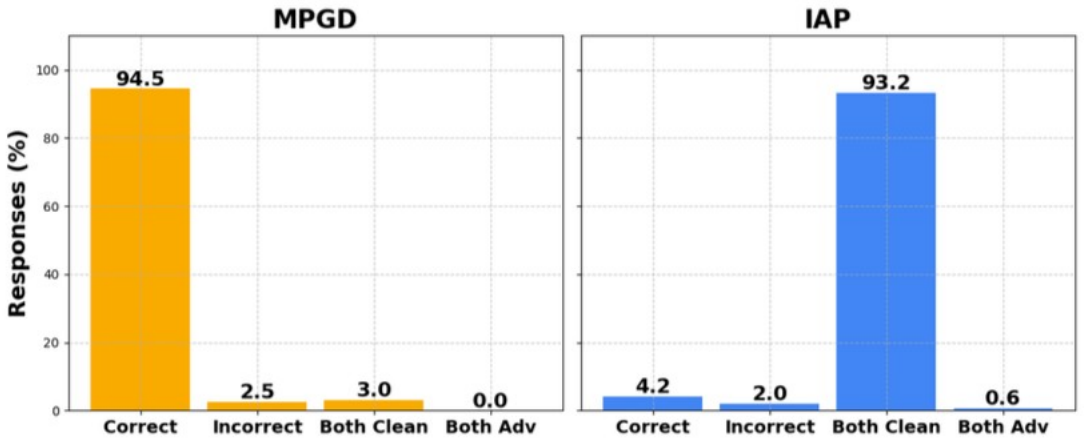


Figure 4. Human perceptability study. “Correct” means correct selection, and “Both Clean” means considering both images clean.

Experiments

- Attack Stealthiness against Defenses
  - IAP: Bypasses all defenses with high success rates (~100% ASR).

Method	Jedi	Jujutsu	SAC	DW	DIFFender	DiffPAD
Google Patch	46.8	0.0	2.7	1.4	35.5	33.2
LaVAN	50.9	0.3	3.8	54.0	53.2	39.8
GDPA	67.1	94.0	7.4	1.3	57.0	52.1
MPGD	68.2	95.1	11.6	79.0	95.7	92.1
IAP	78.6	99.8	100	89.8	99.8	98.6

Table 2. Comparisons of ASR (%) between different attack methods against various patch defenses.

Experiments

- Ablation Study
  - w/o Update Rule:
    - High noise, visually obvious (SSIM: 0.06).
  - w/o Loss Term:
    - Poor blending, distinct edges (SSIM: 0.78).
  - w/o Localization:
    - Patch covers salient features (e.g., Rat's face).
  - Ours (IAP):
    - Seamlessly blends into background texture (SSIM: 0.94).

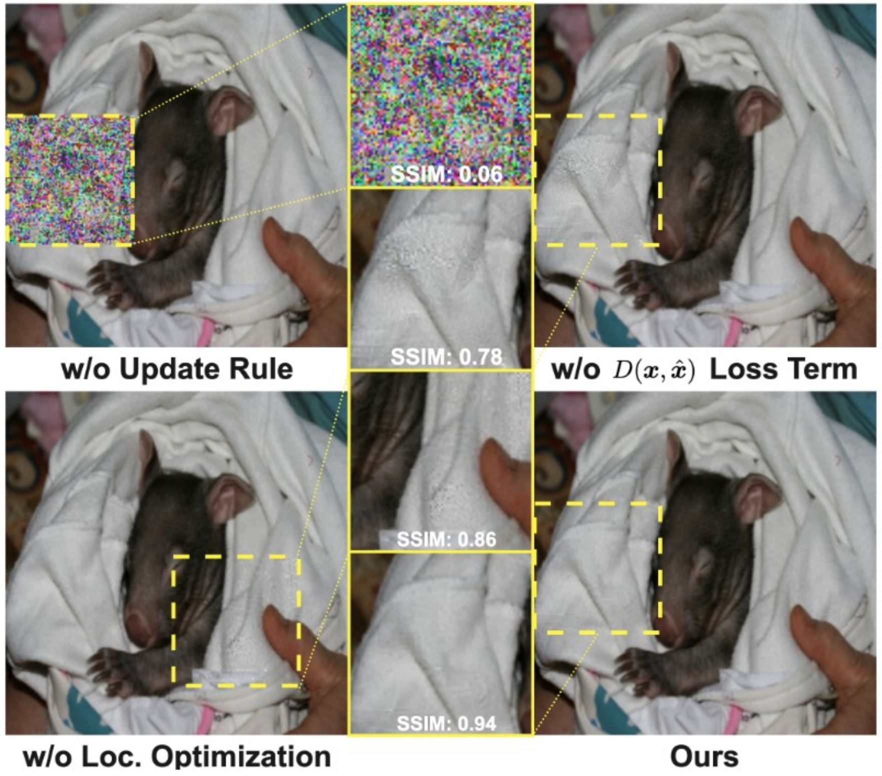


Figure 6. Ablation study on the impact of IAP's components.

Experiments

- black-box attack scenarios
  - verify that the more similar the architecture of the model, the better the attack works

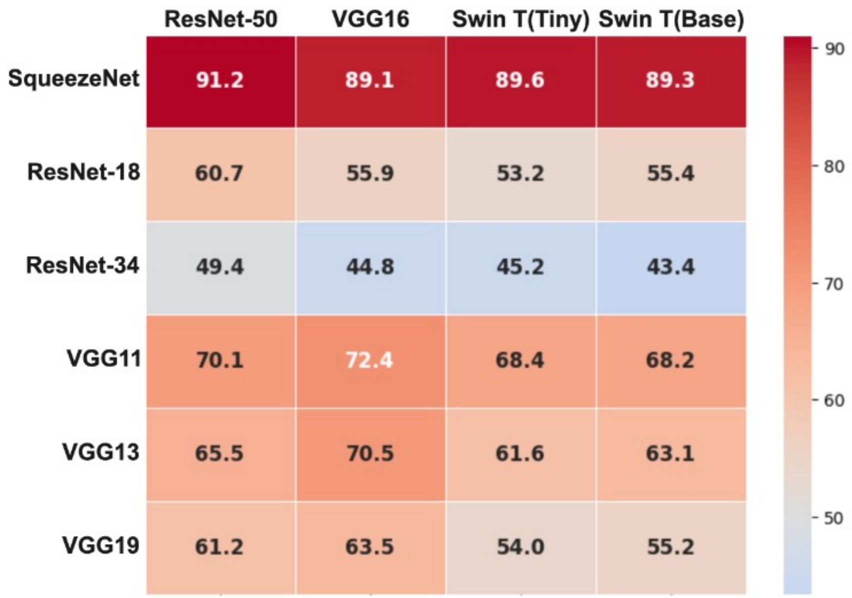


Figure 7. Transferability of IAP measured by ASR (%) on ImageNet. The first row represents the substitute model, and the first column represents the target models.

Conclusion

- IAP : designed to generate imperceptible adversarial patches.
- high stealth , targeted attack efficacy
- IAP also showed promise, both in black-box transferability and in the physical attack domain
- It introduces additional computational overhead (calculate G)
    - single NVIDIA A100 GPU -> iteration:379 , 19sec