结合 **演化搜索（Evolutionary Search）** 与 **梯度对齐（Gradient Alignment）** 的新型对抗攻击想法。我们将这种方法命名为 **EATA (Evolutionary Aligned Transformation Attack)**。

演化对齐变换攻击 (EATA) 技术方案

1. 背景与动机

目前的输入变换攻击（如 BSR 1111、DIM ）主要依赖随机采样。虽然 BSR 通过打乱图像内在关系（Intrinsic Relation）来破坏注意力热图（Attention Heatmaps）并提升迁移性，但均匀随机采样会导致梯度方差过大，且包含大量对攻击贡献较小的"冗余变换"。

EATA 的核心思想：

- **演化搜索**：在输入变换的参数空间内进行在线优化，实时寻找最能激发模型 Loss 增长的"最优变换"组合。
- **梯度对齐**：通过度量梯度方向的一致性，滤除噪声梯度，保留具有强迁移性的核心梯度方向。

---

2. 算法原理

2.1 变换参数空间定义

定义 BSR 变换为 $\mathcal{T}(x;\pi)$，其中参数 $\pi = \{S, \mathrm{B}\}$：

- S：块打乱的置换矩阵（Permutation Matrix）
- $\mathrm{B} = \{\beta_1, \ldots, \beta_{n^2}\}$：每个块独立的旋转角度

2.2 演化搜索策略 (Evolutionary Strategy)

我们不再直接计算梯度的平均值 6666，而是先在推理侧寻找最优参数集 Π^*。

目标函数：

$$\pi^* = \arg\max_{\pi} J(f(\mathcal{T}(x^{adv};\pi)), y; \theta)$$

其中 J 是交叉熵损失 7，f 是替代模型 8。通过对参数种群进行**变异 (Mutation)** 和 **选择 (Selection)**，我们在不计算梯度的情况下筛选出最有效的变换形态。

2.3 梯度对齐与加权 (Gradient Alignment)

为了增强迁移性，我们借鉴了多目标优化的思想。若多个变换下的梯度方向高度一致，则该方向更有可能代表了不同模型间的"通用决策边界破坏方向"。

定义第 k 个优选变换生成的梯度为 $g_{\pi_k} = \nabla_{x^{adv}} J(\mathcal{T}(x^{adv}; \pi_k), y)$。

最终的综合梯度 $\bar{g}$ 计算如下：

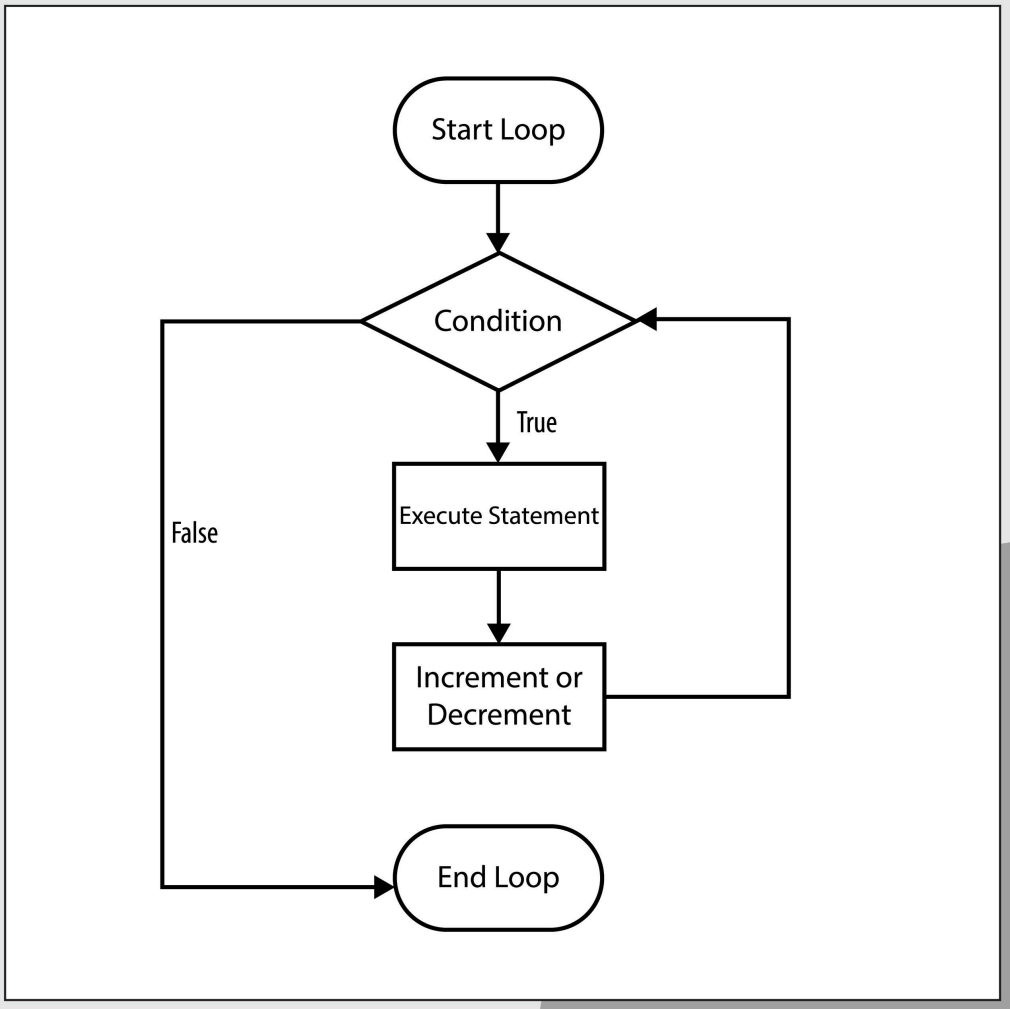
$$\bar{g} = \sum_{k=1}^{K} w_k \cdot g_{\pi_k}$$

其中权重 w_k 由当前梯度与全局动量 m_t 的余弦相似度决定：

$$w_k = \text{Softmax}\left(\frac{g_{\pi_k} \cdot m_{t-1}}{\|g_{\pi_k}\|_2 \cdot \|m_{t-1}\|_2}\right)$$

---

3. 算法流程 (Algorithm)



算法 1：EATA 迭代攻击

输入：替代模型 f，Loss 函数 J，原图 x，扰动限制 ϵ，迭代次数 T，种群规模 M，存活规模 K。

输出：对抗样本 x^{adv}。

1. 初始化 $x_0^{adv} = x, m_0 = 0, \alpha = \epsilon/T$
2. **For** $t = 0$ **to** $T - 1$:
   - **A. 采样与评估**:
     - 随机生成 M 组参数 $\{\pi_1, \ldots, \pi_M\}$。
     - 计算推理损失 $L_m = J(f(\mathcal{T}(x_t^{adv}; \pi_m)), y)$。
   - **B. 演化更新**:
     - 选取 L_m 最高的 K 组参数作为“精英种群”。
     - 对精英种群施加微小扰动（变异），替换较差的 $M - K$ 组。
     - 重新评估并锁定最终的优胜变换集 Π_{best}。
   - **C. 梯度计算与对齐**:

- 计算 Π_{best} 中每个变换对应的梯度 g_{π_k}
- 计算对齐权重 w_k（基于余弦相似度）。
- 聚合梯度 $\bar{g}_t = \sum w_k g_{\pi_k}$。

- **D. 扰动更新**：
  - 更新动量 $m_{t+1} = \mu \cdot m_t + \frac{\bar{g}_t}{\|\bar{g}_t\|_1}$
  - 更新样本 $x_{t+1}^{adv} = \text{Clip}(x_t^{adv} + \alpha \cdot \text{sign}(m_{t+1}))$

3. **End For**
4. **Return** x_T^{adv}

---

4. 理论优势分析

1. **动态适应性**：相比于 BSR 的固定随机分布，EATA 能够针对每一张特定的图像，在搜索空间内找到最能打破该图像语义结构（Semantic Relation）的变换方式
2. **噪声抑制**：通过梯度对齐机制，算法自动过滤了那些仅对当前白盒模型有效、但在变换空间中表现不稳定的“过拟合方向”
3. **计算效率优化**：由于演化搜索阶段仅涉及前向推理（Inference），其计算开销远小于反向传播。在总梯度计算次数受限的情况下，EATA 能通过更精准的梯度方向提升攻击成功率。
