LED物理世界对抗性样本

背景

主要贡献

- **提出了面向物理 LED 对抗样本的两阶段黑盒优化框架**：利用 **粒子群优化（PSO）** 在简化空间中筛选关键 LED 子集，再用 **协方差矩阵自适应进化策略（CMA-ES）** 对子集内亮度与 RGB 参数进行连续细调，实现了在高维、非凸、不可微的物理扰动空间中高效搜索。
- **设计了适应物理约束的连续优化方法**：通过在实数域 $[-6, 6]$ 上搜索并使用 sigmoid 映射回 $[0, 1]$，既保持 CMA-ES 的统计自适应性，又确保 LED 参数满足真实物理可控范围。
- **提出了多任务自适应权重调节机制**：在同时攻击检测器与分类器时，引入基于滑动窗口相关性的权重更新方法；当检测与副任务（可见度或分类器概率）出现负相关时自动提升惩罚权重，在协同时且副任务表现足够好时适度下调，实现了对抗效果与副任务约束的动态平衡。
- **实现了仿真—实物闭环验证流程**：基于 Blender 与 3D 重建快速搭建虚拟场景，得到最优 LED 配置后映射到真实硬件（WS2812B 阵列 + 控制器），在实际相机采集条件下验证了仿真结果的可迁移性与攻击有效性。

Methods

设有M个可控 LED（编号 $j = 1, \ldots, M$），每个 LED 的物理可控参数为亮度与 RGB 三通道，共 4 个连续量：

$$x_j = [b_j, r_j, g_j, b_j]^\top \in [0, 1]^4, \tag{5}$$

其中 b_j 表示亮度（brightness），r_j, g_j, b_j 三个为颜色通道。整个系统的参数向量记为

$$\mathbf{x} = [x_1^\top, x_2^\top, \ldots, x_M^\top]^\top \in [0,1]^{4M}. \tag{6}$$

受攻击模型为黑盒，只能查询模型输出（例如检测器的平均置信度或分类器的 Top-1 置信度）。定义适应度（目标）函数 $F(\mathbf{x})$ 为我们要最小化的指标（例如目标检测平均置信度的加权和或分类 Top-1 置信度），即：

$$\min_{\mathbf{x}\in[0,1]^{4M}} F(\mathbf{x}). \tag{7}$$

为提高搜索效率，我们将优化分为两阶段：先用 PSO 在简化空间搜索“要点亮哪些 LED（子集）”，再用 CMA-ES 在所选子集上做连续参数细调。

阶段一 — Coarse：PSO 子集搜索（子集选择）

目标：在低维布尔空间中选出一个影响力大的 LED 子集 $S \subseteq \{1, \ldots, M\}$。

1.1 问题简化

为降低维数，我们将每个 LED 的颜色与亮度固定为一个预设值（例如紫色 $x^{(0)}$），仅优化每个 LED 的开关状态 $s_j \in \{0,1\}$。用向量 $\mathbf{s} = [s_1, \ldots, s_M]$。

在实现上，可以用连续编码 $y_j \in \mathbb{R}$（或 $[0,1]$）并以阈值方式得到二值：

$$s_j = \mathbb{I}[y_j > \tau_{\text{th}}], \quad \tau_{\text{th}} = 0.5. \tag{8}$$

1.2 Ibest PSO

我们采用 l-best PSO（局部邻域拓扑）以鼓励多样性，速度/位置更新为经典形式（使用 constriction 因子 κ）：

$$\mathbf{v}_i^{t+1} = \kappa\big(\mathbf{v}_i^t + c_1 r_1(\mathbf{p}_i^t - \mathbf{x}_i^t) + c_2 r_2(\ell_i^t - \mathbf{x}_i^t)\big), \tag{9}$$

其中 i 表示粒子索引，$\mathbf{p}_i^t$ 为粒子历史个人最优（pbest），ℓ_i^t 为其邻域最优（lbest），$r_1, r_2 \overset{\text{iid}}{\sim} \mathcal{U}(0,1)$。

1.3离散化与子集输出

粒子位置 $\mathbf{x}_i$（维度为 M，代表开/关概率）在每次评估时通过阈值/采样映射为二值 $\mathbf{s}$，并以固定颜色 $x^{(0)}$ 组合还原为物理参数用于仿真或真实评估。最终输出候选子集集合（取若干 top-k 粒子最优子集）供 Finetune 阶段逐一细调。

阶段二 — Finetune：CMA-ES 连续参数细调

目标：对候选子集 S 中的 LED 做连续参数 $\in [0,1]^{4|S|}$ 的精调，最小化 F。

2.1 实数域搜索与 Sigmoid 映射

为不直接破坏 CMA-ES 的自适应统计特性，在内部在无界实数域 $y \in \mathbb{R}^d$（$d = 4|S|$）中搜索，并通过 sigmoid 映射到物理区间：

$$x = \sigma(y) \equiv \frac{1}{1+\exp(-y)} \in (0,1). \tag{10}$$

为让搜索覆盖近 $[0,1]$ 边界，可以在 y 空间使用较大范围（例如初始均值可设在 [-1,1]]，或更宽 ([-3,3]]），并通过缩放使映射分布合适。实现细节：在文中我们建议搜索区间 ([-6,6]（经验值），以便 sigmoid 可实现接近 $[0,1]$的值。

2.2 CMA-ES 基本步骤（简记）

采用标准 CMA-ES 采样与更新，核心为：

- 采样：

$$\mathbf{y}_i^{t+1} \sim \mathcal{N}(\mathbf{m}^t, (\sigma^t)^2 \mathbf{C}^t), \quad i = 1, \ldots, \lambda, \tag{11}$$

- 选择并更新均值：

$$\mathbf{m}^{t+1} = \sum_{k=1}^{\mu} w_k \mathbf{y}_{k:\lambda}^{t+1}, \tag{12}$$

- 协方差矩阵与步长更新（rank-one 和 rank-μ 更新，CSA）：

$$\mathbf{C}^{t+1} = (1 - c_1 - c_\mu)\mathbf{C}^t + c_1 \mathbf{p}_c \mathbf{p}_c^\top + c_\mu \sum_{k=1}^{\mu} w_k \mathbf{y}_{k:\lambda} \mathbf{y}_{k:\lambda}^\top, \tag{13}$$

（以上变量符号与 Hansen 标准定义一致；）

2.3 k-random-start（多次重启）

为了增强稳定性并减少对初值敏感性，对同一候选子集 S 进行 k 次独立 CMA-ES 运行（不同初始 $\mathbf{m}^0$、初始步长 σ^0 或随机种子），取 k 次中最优结果：

$$\hat{\mathbf{x}} = \arg\min_{i=1,\dots,k} F\big(\sigma(\mathbf{y}_{\text{best}}^{(i)})\big). \tag{14}$$

$k = 3 \sim 7$（经验），默认k=3。

2.4 适应度函数设计

最终的优化目标函数由三部分组成：

(a) 检测器抑制损失

对于图像 I，检测器返回一组阈后框 $\mathcal{D}_\tau(I)$。优化目标是阈后置信度的总和：

$$L_{\text{det}}(I;\tau) = \sum_{i \in \mathcal{D}_\tau(I)} s_i, \tag{15}$$

若无检测框，则 $L_{\text{det}} = 0$。

(b) 可见度约束

利用与原图 I_0 的像素均方根差：

$$L_{\text{vis}}(I, I_0) = \sqrt{\frac{1}{|\Omega|}\sum_{p\in\Omega} \|I(p) - I_0(p)\|_2^2}, \quad \lambda_{\text{pix}} = 0.04. \tag{16}$$

(c) 分类器损失

压低原始图像的 Top-1 类别 $\hat{y}_0$ 在当前图上的概率：

$$L_{\text{cls}}(I) = P(\hat{y}_0 \mid I), \tag{17}$$

若分类器不可用则置零。

2.5 最终适应度函数设计

固定权重版本

早期实验中，我们使用固定权重组合：

$$L = \alpha\, L_{\text{det}} + \beta\, \lambda_{\text{pix}} L_{\text{vis}} + \gamma\, L_{\text{cls}}, \tag{18}$$

其中 $\alpha, \beta, \gamma \geq 0$ 为常数。

自适应权重版本

损失定义为：

$$L^{(t)} = w_1^{(t)} L_{\text{det}}^{(t)} + w_2^{(t)} \lambda_{\text{pix}} L_{\text{vis}}^{(t)} + w_3^{(t)} L_{\text{cls}}^{(t)}, \qquad \sum_{i=1}^{3} w_i^{(t)} = S^*, \tag{19}$$

其中：

- L_{det}：检测器抑制损失；
- L_{vis}：像素 L2 距离正则，$\lambda_{\text{pix}} = 0.04$；
- L_{cls}：分类器概率损失；
- $\mathbf{w}^{(t)} = (w_1^{(t)}, w_2^{(t)}, w_3^{(t)})$，保持和为常数 S^*（实现中 `sum_target`）。

维护长度为 W 的滑动窗口：

$$\mathcal{H}_{\text{det}} = \{L_{\text{det}}^{(t-W+1)}, \ldots, L_{\text{det}}^{(t)}\}, \quad \mathcal{H}_{\text{vis}}, \quad \mathcal{H}_{\text{cls}}. \tag{20}$$

计算一阶差分：

$$\Delta\mathcal{H}_k = \{L_k^{(\tau)} - L_k^{(\tau-1)}\}_{\tau=t-W+2}^{t}, \quad k \in \{\text{det}, \text{vis}, \text{cls}\}. \tag{21}$$

得到相关性系数：

$$r_{\text{det,vis}} = \text{corr}\big(\Delta\mathcal{H}_{\text{det}}, \Delta\mathcal{H}_{\text{vis}}\big), \qquad r_{\text{det,cls}} = \text{corr}\big(\Delta\mathcal{H}_{\text{det}}, \Delta\mathcal{H}_{\text{cls}}\big). \tag{22}$$

权重更新规则

设相关阈值为 ρ，上调率 $\eta_{\uparrow}$，下调率 $\eta_{\downarrow}$，目标阈值分别为 $T_{\text{vis}}, T_{\text{cls}}$。更新规则如下：

1.**冲突时上调惩罚权重**（负相关）：

$$\text{若 } r_{\text{det,vis}} < -\rho: \quad w_2 \leftarrow w_2(1+\eta_{\uparrow}), \tag{23}$$

2.**协同时且惩罚均值不高则下调**：

$$\text{若 } r_{\text{det,vis}} > \rho \ \wedge \ \overline{L}_{\text{vis}} < T_{\text{vis}}: \quad w_2 \leftarrow w_2(1-\eta_{\downarrow}), \tag{24}$$

> 如果副任务的均值本来就很低（比如 L2 距离已经小于设定的目标 T_{vis}，或者分类器概率已经低于 T_{cls}），说明副任务已经“够好”了，不需要再过度约束。在这种情况下，才允许下调权重。
>
> **反例**：假设相关性 > 0，但可见度均值其实很高（扰动很明显），那就不能因为“它和检测器损失协同”就盲目降低权重，否则会导致扰动越来越明显，破坏物理隐蔽性。
>
> **T_vis,T_cls 是相当于保证**只有在副任务表现已经足够好（低于阈值）时，才允许去下调其权重；否则即使是r > 0，也要保持或提高权重，避免副任务恶化。

3.**主任务收敛缓慢时微调检测器权重**：

若

$$\overline{\Delta L_{\text{det}}} > -\varepsilon, \quad (\overline{L}_{\text{vis}} \geq T_{\text{vis}} \ \vee \ \overline{L}_{\text{cls}} \geq T_{\text{cls}}), \tag{25}$$

则

$$w_1 \leftarrow w_1(1+0.05). \tag{26}$$

4.**裁剪与归一化**：

$$w_i \leftarrow \text{clip}(w_i;\ w_i^{\min}, w_i^{\max}), \quad i=1,2,3, \tag{27}$$

Early Stopping

若检测器返回 success $=\text{True}$ 或 $L_{\text{det}} \leq \tau_{\text{det}}$（阈值由 `CLASS_LOSS_THRESHOLD` 设定），立即触发早停标记，结束当前运行。

实验

状态	实验名称	实验目的	主要对比/设置	指标输出
✅	基础攻击有效性	验证 LED 两阶段攻击能否显著降低模型性能	YOLOv8 (det)，ResNet-101(cls)，仿真场景	confidence、Top-1 prob、L2 距离
✅	仿真→实物迁移	检验在真实 LED 阵列上是否仍有效	Blender 优化结果映射到真实硬件	攻击成功率、置信度下降、迁移率
	鲁棒性测试	考察在不同物理条件下的稳定性	光照强度、相机距离/角度、噪声/模糊	指标分布（均值 ±SD）
	多任务冲突 & 自适应权重	展示自适应权重缓解 det/cls 冲突的效果	固定权重 vs 自适应权重	sum_conf、Top-1 prob、相关系数轨迹、权重曲线
	消融实验	验证设计的必要性	(a) 两阶段 vs 单阶段，(b) Sigmoid vs Clamp，(c) k-start 数量?	成功率、收敛速度、L2
	可感知性分析?	证明扰动的隐蔽性?	L2、SSIM、LED 点亮率	视觉差异图、定量指标
	跨模型迁移性（看来得及吗？）	考察同一配置对其它模型是否有效	ResNet-101→ ResNet-50；YOLOv8 → 其它检测器	降幅对比、迁移率