Algorithm 1: Zeroth-Order Optimization of Natural Light attack

Input: Victim classifier f, Input image x, Ground truth y, Light generator $\mathcal{G}$, Update times t_{max}, Distribution of parameter P, Step size of estimate gradient δ, Step size of optimize γ, Restart threshold τ

Output: Adversarial image x^*_{adv} with maxi. score

$x^*_{adv} \leftarrow x$
$score^* \leftarrow \mathcal{L}_{CE}(y, f(x^*_{adv}))$
// Optimize k rounds
for $i \leftarrow 1$ *to* k **do**
  sample $\mathcal{P}$ *from* P
  // Optimize each $\mathcal{P}$ t_{max} *times*
  for $t \leftarrow 1$ *to* t_{max} **do**
    $x_{adv} \leftarrow x + \mathcal{G}(x, \mathcal{M}_{\mathcal{P}})$
    $score \leftarrow \mathcal{L}_{CE}(y, f(x_{adv}))$
    if $score > score^*$ **then**
      $x^*_{adv} \leftarrow x_{adv}$
      $score^* \leftarrow score$
    // Estimate gradient of each parameter in $\mathcal{P}$
    for $\mathcal{P}_j$ *in* $\mathcal{P}$ **do**
      $\mathcal{P}_j \leftarrow \mathcal{P}_j + \delta$
      $x^{(j)}_{adv} \leftarrow x + \mathcal{G}(x, \mathcal{M}_{\mathcal{P}})$
      $score^{(j)} \leftarrow \mathcal{L}_{CE}(y, f(x^{(j)}_{adv}))$
      $grad_j \leftarrow (score^{(j)} - score)/\delta$
      if $score > score^*$ **then**
        $x^*_{adv} \leftarrow x^{(j)}_{adv}$
        $score^* \leftarrow score^{(j)}$
      $\mathcal{P}_j \leftarrow \mathcal{P}_j - \delta$
    // Random start with vanish gradient
    if $||grad||_2 < \tau$ **then**
      sample $\mathcal{P}$ *from* P
    // Update parameter
    else
      $\mathcal{P} \leftarrow \mathcal{P} + \gamma \frac{grad}{||grad||_2}$

‘