先根据最小图像特征选出生成对抗时要修改的点：

最小图像特征：给定一个图像x和其对应的分类y，我们将简化这个图像变为x',使得x'还被高置信度的分类到y。

由于简化后的图像被认定为x图像的最小图像特征，改变这个特征区域可显著改变分类结果。

1. 先根据图像边界分割算法将图像分割成一个个小区域。

$$r(r_1, r_2, \dots, r_n)$$

其中n为图像所分成的区域个数。

2. 每次迭代去除使置信度下降最低的一个区域，直到分类发生错误。

$$r' = min\ r \quad s.t. \quad F(r') = F(r)$$

其中F为分类器。

对于选定的最小图像特征区域进行对抗样本生成

1. 对最小图像特征区域r'的扰动δ进行按像素编码：

$$\delta = (\delta_1, \delta_2, \dots, \delta_m)$$

其中m为r'所有的像素点个数。

2. 使用遗传算法生成对抗样本：

$$min \ ||\delta||^2 \ s.t. \quad F(\mathrm{x}+\delta) = \mathrm{t}$$

其中x为源图像，t为对抗样本的目标分类。

（注：其中必然要给出限制

$$||\delta||^2 \leq \mathrm{K}$$

，其中K为扰动上限。）

为了使得更方便使用遗传算法找到最优解，将上述目标转换成如下目标公式：

$$\arg_\delta \ \min \ f'(x+\delta) + c||\delta||^2$$

其中，c为比重参数；$f'(x) = \Sigma_i^{i \neq t} F_i(x) - F_t(x)$，如果要对抗蒸馏网络，$f'(x)$需要被设计为$f'(x) = \Sigma_i^{i \neq t} Z_i(x) - Z_t(x)$。其中Z为神经网络输出的前一层，也就是logit层。