结合高斯滤波与MASK的G-MASK人脸对抗攻击

李倩，向海昀*，张玉婷，甘昀，廖浩德
（西南石油大学计算机科学学院，四川 成都 610500）

摘　要：深度神经网络的快速发展使其在计算机视觉和自然语言处理等领域取得较大成功，但是对抗攻击会导致神经网络的表现性能降低，对各类系统的安全保密性造成严重威胁。现有黑盒攻击方法在人脸识别中性能表现较差，攻击成功率较低且生成对抗样本迁移性不高。为此，提出一种结合高斯滤波与掩码的对抗攻击方法G-MASK。利用Grad-CAM输出的热力图确定对抗样本的掩码区域，使其只在掩码区域施加扰动，提高黑盒攻击成功率，采用扰动集成方法提高黑盒迁移能力，增强黑盒攻击鲁棒性，对生成的扰动进行高斯平滑处理，降低集成模型之间干扰噪声的差异，提高图像质量且增强扰动掩蔽性。实验结果表明，针对不同的人脸识别模型，G-MASK方法在保证白盒攻击成功率较高的条件下能够显著提升黑盒攻击效果，并具有更优的掩蔽性，经过模型扰动集成的对抗样本白盒攻击成功率均提高至98.5%以上，黑盒攻击成功率最高达到75.9%，与快速梯度符号法（FGSM）、迭代快速梯度符号法（I-FGSM）和动量迭代梯度符号法（MI-FGSM）相比分别平均提升12.1、10.6和8.2个百分点，充分验证了该方法的有效性。

关键词：对抗样本；黑盒攻击；人脸识别；高斯滤波；掩码
源代码链接：https://github.com/nikkiicn/G_MASK.git
中图分类号：TP181　　**文献标志码：**A　　**DOI：**10.19678/j.issn.1000-3428.0066455

G-MASK Facial Adversarial Attack Combining Gaussian Filtering and MASK

LI Qian，XIANG Haiyun*，ZHANG Yuting，GAN Yun，LIAO Haode
（School of Computer Science，Southwest Petroleum University，Chengdu 610500，Sichuan，China）

【Abstract】The rapid development of deep neural networks has led to significant success in fields such as computer vision and natural language processing. However，adversarial attacks may inhibit the performance of neural networks，posing a serious threat to the security and confidentiality of various systems. Existing black-box attack methods perform poorly in facial recognition，with a low success rate and low transferability of generated adversarial samples. To this end，a G-MASK adversarial attack method combining Gaussian filtering and mask is proposed. Using the heat map output by Grad-CAM to determine the mask area of adversarial samples，the mask area is perturbed to improve the success rate of black-box attacks. The perturbation integration method is used to improve the black-box migration ability and enhance attack robustness. Gaussian smoothing is applied to the generated perturbations to reduce the difference in interference noise between integrated models，improve image quality，and enhance disturbance masking. Experimental results show that for different facial recognition models，the G-MASK method significantly improves the effectiveness of black-box attacks while ensuring a high success rate of white-box attacks and a better masking ability. Following model perturbation integration，the success rate of white-box attacks on adversarial samples exceeds 98.5%，while the success rate of black-box attacks reaches 75.9%，which is consistent with the fast gradient sign method. Compared with Fast Gradient Symbolic Method（FGSM），Iteration-Fast Gradient Symbolic Method（I-FGSM），Moenttum Iteration-Fast Gradient Symbolic Method（MI-FGSM）yields average improvements of 12.1，10.6，and 8.2 percentage points.

【Key words】adversarial sample；black-box attack；facial recognition；Gaussian filtering；mask

0　引言

深度学习理论的不断发展使得基于该技术的人工智能在各个行业领域中崭露头角，如图像分类[1]、计算机视觉[2]、自然语言处理[3]、人脸识别[4]等。同时，随着对神经网络的不断深入研究，结构更为复杂的网络模型（如AlexNet[5]、VGGNet[6]、InceptionNet[7]、ResNet[8]等）相继面世，它们的出现不仅能够提升预测结果的精确度，还能够满足更多的场景需求。

但是随着对抗攻击的出现，导致深度神经网络模型分类错误，为使用深度神经网络的人工智能系统带来极大的安全风险。对于人脸识别系统，攻击者通过在原图像上添加细微的扰动生成人脸对抗样本，导致人脸系统识别错误。文献[9]提出对眼镜进行对抗性干扰的方法，使得佩戴眼镜的攻击者被错误识别。现有人脸识别攻击大多是通过白盒实现

收稿日期：2022-12-07　　**修回日期：**2023-03-23
基金项目：国家自然科学基金青年科学基金（61503312）。
通信作者E-mail：*xhy11202021@163.com

的，但是在现实世界中，攻击者一般无法直接访问模型的详细信息。因此，黑盒攻击被进一步提出，它是利用白盒攻击生成对抗样本的迁移能力实现的。但是由于不同模型的结构和参数存在差异，因此使得对抗样本出现“过拟合”[10]，因对抗样本迁移能力较差，导致黑盒攻击成功率降低。

为提高对抗样本的迁移性，本文提出一种集成的对抗攻击方法 G-MASK，将基于梯度加权的类激活映射方法运用于算法中生成掩码区域，以确定扰动添加的最佳区域，使得只需极少的扰动就能实现有效的攻击，该方法能够有效增加对抗攻击迁移性，提高攻击成功率。同时对生成的扰动进行高斯平滑处理，使得生成扰动的每个像素都与周围的像素产生相关性，进而降低集成模型之间扰动的差异，增强对抗样本隐蔽性，提升图像质量。

1 相关工作

1.1 对抗样本生成算法

对抗样本是指通过在原始图像中加入人眼难以察觉的细小扰动，使得人眼能正确识别图像但机器识别出错的一类样本。设输入的样本为 $x\in\mathbb{R}^d$，正确的类别分类为 y，$y=f(x)$，添加的扰动为 δ，生成对抗样本 $y^*=f(x+\delta)$。对抗样本的生成过程如式(1)和式(2)所示：

$$x^{\text{adv}}=x+\delta \tag{1}$$

$$\text{argmax}J(x^{\text{adv}},y)\ \text{s.t.}\|\delta\|\leqslant\varepsilon \tag{2}$$

在深度学习领域中，将对抗攻击方法分为 3 类，分别为基于优化、基于梯度、基于生成对抗网络的方法。本文主要对基于梯度的经典攻击方法进行简要叙述。

文献[11]提出的快速梯度符号法(FGSM)是在损失减少的反方向上增加扰动函数，然后不断进行梯度更新来实现对抗样本的生成。对抗样本的生成过程如式(3)所示：

$$x^{\text{adv}}=x+\varepsilon\cdot\text{sign}(\nabla_x L(x,y^{\text{true}};\ \theta)) \tag{3}$$

其中：sign(·)为符号函数；$\nabla_x L(x,y^{\text{true}};\ \theta)$表示模型的梯度。

文献[12]提出的基础迭代方法(I-FGSM)是在快速梯度符号法的基础上进行改进，将原有的单步更新改进为迭代更新，将扰动大小限制在一定大小的邻域内。I-FGSM 对抗样本的生成过程如式(4)所示：

$$x_{n+1}^{\text{adv}}=x_n^{\text{adv}}+\alpha\cdot\text{sign}(\nabla x_n^{\text{adv}}L(x_n^{\text{adv}},y^{\text{true}})) \tag{4}$$

其中：α为步长大小；$x_0^{\text{adv}}=x$；n为迭代次数。

文献[13]提出的动量迭代快速梯度符号方法(MI-FGSM)是在迭代方法基础上进行改进，动量的添加使得模型具有更强的迁移能力。MI-FGSM 对抗样本的生成过程如式(5)和式(6)所示：

$$g_{n+1}=\mu\cdot g_n+\frac{\nabla x_n^{\text{adv}}L(x_n^{\text{adv}},y;\ \theta)}{\|\nabla x_n^{\text{adv}}L(x_n^{\text{adv}},y;\ \theta)\|_1} \tag{5}$$

$$x_{n+1}^{\text{adv}}=x_n^{\text{adv}}+\alpha\cdot\text{sign}(g_{n+1}) \tag{6}$$

其中：g_{n+1}为第 $n+1$ 次的迭代梯度；$\|\cdot\|_1$为 L1 范数。

1.2 集成模型攻击

对抗样本如果能够欺骗多个模型，那么说明它具有较强的迁移能力，可以将其运用于其他黑盒模型中进行攻击。因此，通过对模型进行集成可以实现更好的攻击效果，达到更高程度的迁移目的。文献[14]提出扰动集成方法，将多个白盒模型的扰动在迭代算法基础上实现集成效果，从而侧面提高黑盒攻击成功率。与受到模型输出维度限制的 logits 集成方法相比，扰动集成能够满足黑盒模型的决策边界要求。因此，本文选择使用扰动集成方式进行实验。扰动集成的计算式如式(7)所示：

$$\delta_m=f_\theta(\delta_{m-1},y;\ L_i) \tag{7}$$

其中：δ表示对扰动进行集成；f为模型采用的对抗攻击算法；θ为模型参数；y表示真实标签；$L=\{L_1,L_2,\cdots,L_n\}$为白盒模型集合。

1.3 深度学习的人脸识别

人脸识别(FR)模型是通过检测照片中的人脸，然后计算该人脸与其他人脸的距离对 2 张人脸进行匹配，确定是否属于同 1 张人脸。距离度量一般采用 L2 范数或余弦相似度，将度量方式与阈值一起使用，以计算面部的贴近度。在 DeepFace[15]和 DeepID[16]中，将人脸识别视为 1 个多分类问题，并通过深度神经网络学习经过 Softmax 损失的特征。

本文选用 7 个基于深度学习训练的人脸识别网络模型作为目标模型展开研究：ResNet 50，ResNet 101，ResNet 152[17]，SEResNet 50，SEResNet 101[18]，Attention 56[19]，MobileNet[20]。其中，ResNet 50、ResNet 101、ResNet 152 是 3 个深度逐渐增加的残差网络。SEResNet 50、SEResNet 101 是将 SE Block 嵌入到 ResNet 中构建的网络。

2 方法实现

2.1 Grad-CAM 方法

Grad-CAM[21]是类激活映射方法，它是在梯度加权的基础上实现的。Grad-CAM 用于突出输入图像中与网络预测相关的重要区域，对深度神经网络中的卷积层梯度进行计算，得到不同神经元对样本不同类别的重要程度，将这些区域进行高亮显示，从而生成注意力热图。Grad-CAM 的计算式如式(8)和式(9)所示：

$$a_k^t=\frac{1}{z}\sum_i\sum_j\frac{\partial y^t}{\partial A_i,j^k} \tag{8}$$

$$L^t{}_{\text{Grad-CAM}}=\text{ReLU}\left(\sum_k\alpha_k^tA^k\right) \tag{9}$$

其中：y^t表示 t 类别的 logits；A为特征图；k表示 A 的通道；i,j分别为 A 的横纵坐标；z为 A 的长宽相乘。

基于此，本文尝试在人脸图像上生成注意力区域热力图，实验结果表明，面部区域的颜色更加突出，且五官区域的特征点更明显。人脸关注区域热力图如图1所示。

(a)原图 (b)热力图 (c)叠加图像

图1 人脸关注区域热力图

Fig.1 Heatmap of facial focus area

2.2 高斯滤波

在图像处理概念下将高斯滤波作为低通滤波器滤除低频噪声，使图像变得模糊以达到平滑的效果。本文在对抗样本中加入高斯滤波的作用是让每个产生扰动的像素和它周围的像素产生关联，使得在各个不同模型之间的噪声差异减小，达到较优的迁移性能。二维高斯函数如式(10)所示：

$$G(x,y)=\frac{1}{2\pi\sigma^2}\mathrm{e}^{-\frac{(x-\mu)^2+(y-\mu)^2}{2\sigma^2}} \tag{10}$$

2.3 损失函数

2.3.1 特征损失

在基于某个模型进行攻击时，为提高攻击成功率，则希望攻击前后人脸图像的特征损失增大，即2张人脸图片特征向量的Cosine相似度减小。Cosine相似度越小说明攻击前后2张图像的相似度越小，就越有可能达到想要攻击的目标，攻击成功率越大，具体计算式如式(11)和式(12)所示：

$$\cos(\boldsymbol{f}_{\mathrm{adv}},\boldsymbol{f}_{\mathrm{ori}})=\frac{\boldsymbol{f}_{\mathrm{adv}}\cdot\boldsymbol{f}_{\mathrm{ori}}}{|\boldsymbol{f}_{\mathrm{adv}}|\cdot|\boldsymbol{f}_{\mathrm{ori}}|} \tag{11}$$

$$\cos\_\mathrm{loss}(\boldsymbol{f},y)=\begin{cases}1-\cos(\boldsymbol{f}_{\mathrm{adv}},\boldsymbol{f}_{\mathrm{ori}}), & y=1\\ \max(0,\cos(\boldsymbol{f}_{\mathrm{adv}},\boldsymbol{f}_{\mathrm{ori}})-\mathrm{margin}), & y=-1\end{cases} \tag{12}$$

其中：$\boldsymbol{f}_{\mathrm{ori}}$表示在预处理过程中已经事先提取好的特征向量；$\boldsymbol{f}_{\mathrm{adv}}$为攻击后生成的对抗样本经过特征网络提取出的特征向量。

2.3.2 扰动损失

为减小攻击后的人脸图片与原始人脸图片之间的差异，本文提出扰动损失L2_loss，减少对人脸图像扰动的添加，约束x^{adv}尽可能地接近x，扰动损失计算式如式(13)所示：

$$\mathrm{L2\_loss}=\|x-x^{\mathrm{adv}}\|_2 \tag{13}$$

其中：x为原样本；x^{adv}为对抗样本。

2.3.3 总损失

综上所述，G-MASK算法的总损失函数由特征损失、扰动损失2部分组成。特征损失主要用于减少生成对抗样本与原始图像的相似度，增大攻击成功率。扰动损失主要用于对扰动大小进行约束，保证图片的质量，使人眼无法察觉扰动。具体的总损失计算式如式(14)所示：

$$\begin{aligned}&\min_{\delta}\ \alpha\cdot\cos\_\mathrm{loss}+\beta\cdot\mathrm{L2\_loss}\\&\mathrm{s.t.}\ x+\delta\in[0,1]\end{aligned} \tag{14}$$

其中：α、β分别为特征损失和扰动损失的权重，用于平衡各项损失。

2.4 G-MASK对抗攻击方法

为保证在扰动添加最少的条件下实现最大的攻击成功率，本文提出G-MASK算法，以提高生成对抗样本的迁移能力。G-MASK算法具体架构如图2所示。

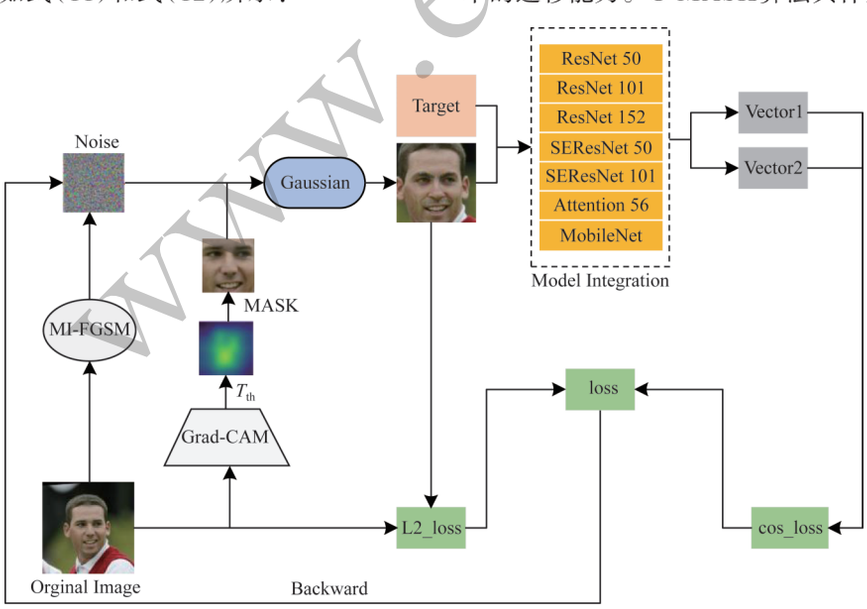

图2 G-MASK人脸对抗攻击算法框架

Fig.2 Framework of G-MASK facial adversarial attack algorithm

首先利用Grad-CAM算法得到输出样本的有效攻击区域，并将其作为掩码MASK；然后将其与通过MI-FGSM[13]算法得到的干扰信息进行叠加，对人脸识别过程中的总损失进行约束；最后将所得到的噪声图进行高斯平滑得到最终的对抗样本。Grad-CAM算法在掩码区域中的具体计算式如式(15)和式(16)所示：

$$x^{*}=x+M_{\mathrm{MASK}}(I_{\mathrm{Grad-CAM}},T_{\mathrm{th}})\cdot\delta \tag{15}$$

$$M_{\mathrm{MASK}}(I_{\mathrm{Grad-CAM}},T_{\mathrm{th}})=\begin{cases}I_{\mathrm{Grad-CAM}},I_{\mathrm{Grad-CAM}}>T_{\mathrm{th}}\\0,\text{其他}\end{cases} \tag{16}$$

其中：T_{th}为施加在掩码上的阈值；$I_{\mathrm{Grad-CAM}}$为利用Grad-CAM得到的输出样本热力图。

为进一步提高对抗样本的迁移性，同时更充分利用目标模型的信息，本文提出使用扰动集成来进行模型集成。受通用对抗扰动[22]的启发，扰动集成的实现方式是在迭代梯度的条件下将白盒模型的扰动进行叠加，使得攻击成功率最大，黑盒迁移能力得到提升。G-MASK单模型算法和扰动集成模型算法如算法1和算法2所示。

算法1 G-MASK(单模型)

输入 图像干净样本$x\in\mathbb{R}^{m}$，网络模型Model，正确分类标签y，攻击区域mask，迭代轮数T，对抗攻击算法f，扰动大小ε，高斯核G，动量因子μ，特征权重α，损失权重β

输出 对抗样本x^{adv}

$x_0=x;t=0;x_0^{\mathrm{adv}}=x;\mathrm{mask}_0=0$//初始化

while t=0 to T−1do

for j=0 to x_m do

$\mathrm{mask}_j=I_{\mathrm{Grad-CAM}}(x_j)$//计算攻击区域

$f=\mathrm{MI-FGSM}(\mathrm{mask}_j)$//MI-FGSM攻击

$\mathrm{new}_{x_j}=x_j+G(f)$//高斯平滑

$y_{i,j}=\mathrm{Model}(x_j)$

$y_{i,j}^{*}=\mathrm{Model}(\mathrm{new\_}x_j)$

$\mathrm{cos\_loss}=\mathrm{CosEmbeddingLoss}(y_{i,j}^{*},y_{i,j},-1)$

$\mathrm{L2\_loss}=\mathrm{L2\_loss}(\mathrm{new}_{x_j},x_j)$//扰动损失

$\mathrm{loss}=\alpha*\mathrm{cos\_loss}+\beta*\mathrm{L2\_loss}$//总损失

通过反向传播求loss

更新mask_j

end for

end while

return $x^{\mathrm{adv}}=x_T^{\mathrm{adv}}$//返回对抗样本

算法2 G-MASK(扰动集成模型)

输入 n个模型n_Models的损失函数$\mathrm{loss}(L_1,L_2,\cdots,L_n)$，图像干净样本$x\in\mathbb{R}^{m}$，正确分类标签$y$，攻击区域mask，迭代轮数$T$，对抗攻击算法$f$及对应参数$\theta$，扰动大小$\varepsilon$，高斯核$G$，动量因子$\mu$，特征权重$\alpha$，损失权重$\beta$

输出 对抗样本x^{adv}

初始化$x_0=x;t=0;x_0^{\mathrm{adv}}=x;\mathrm{mask}_0=0$

while t=0 to T−1 do

for i=0 to n_Models do//n个模型集成

for j=0 to x_m do

$\delta_m=f(\delta_{m-1},y;L_i)$//扰动集成

$\mathrm{mask}_j=I_{\mathrm{Grad-CAM}}(x_j)$//计算攻击区域

$f=\mathrm{MI-FGSM}(\mathrm{mask}_j)$//MI-FGSM攻击

$\mathrm{new}_{x_j}=x_j+G(f)$//高斯平滑

$y_{i,j}=\mathrm{Mean}(\mathrm{Model}(x_j))$

$y_{i,j}^{*}=\mathrm{Model}_i(\mathrm{new\_}x_j)$

$\mathrm{cos\_loss}=\mathrm{CosEmbeddingLoss}(y_{i,j}^{*},y_{i,j},-1)$

$\mathrm{L2\_loss}=\mathrm{L2\_loss}(\mathrm{new}_{x_j},x_j)$//扰动损失

$\mathrm{loss}=\alpha*\mathrm{cos\_loss}+\beta*\mathrm{L2\_loss}$//总损失

通过反向传播求loss

更新mask_j

end for

end for

end while

return $x^{\mathrm{adv}}=x_T^{\mathrm{adv}}$//返回对抗样本

3 实验结果与分析

3.1 实验设置

3.1.1 数据集

本文实验主要在LFW(Labeled Faces in the Wild)[23]数据集上进行，该数据集约有13 000张人脸图片，每张人脸图片对应1个姓名标签，其中有些人脸拥有不止1张图片。该数据集中人脸图片的尺寸是固定的，都是250×250像素。

3.1.2 攻击模型与基线

为证明该方法的有效性，本文使用ResNet 50、ResNet 101、ResNet 152、SEResNet 50、SEResNet 101、Attention 56、MobileNet 7个正常训练得到的模型。

为更好地评估本文算法的有效性，将该算法与3种基线算法(FGSM[11]、I-FGSM[12]和MI-FGSM[13])进行比较，这3种算法都是经典的对抗攻击算法。

3.1.3 参数设置

本文的算法环境在PyTorch 1.10.2框架上设置，显卡配置为NVIDIA GeForce RTX 3050 Ti 4 GB，batch_size为4，使用Adam[24]优化方法在标准的LFW人脸数据集上进行实验。为保证实验结果的客观性，3种基线算法涉及到的参数按照原文献设置。

3.1.4 评价指标

攻击成功率(ASR)是指图像被错误分类的概率。本文方法是一种无目标黑盒攻击，因此只要生成的对抗样本与原图分类出错即可认为攻击成功，

将成功攻击的样本数与总样本数进行比较得到攻击成功率。相应的ASR(计算中用A_{ASR})计算式如式(17)所示:

$$A_{ASR}=\frac{样本成功攻击个数}{样本总个数}\times 100\%=1-\frac{成功分类样本个数}{样本总个数}\times 100\% \tag{17}$$

本文采用峰值信噪比(PSNR,计算中用P_{PSNR})和均方误差(MSE,计算中用M_{MSE})作为图像质量的检测指标,具体计算式如式(18)和式(19)所示:

$$M_{MSE}=\frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\left[I(i,j)-K(i,j)\right]^2 \tag{18}$$

$$P_{PSNR}=10\cdot\lg\left(\frac{MAX_I^2}{M_{MSE}}\right)=20\cdot\lg\left(\frac{255}{\sqrt{M_{MSE}}}\right) \tag{19}$$

其中:m、n表示图片的大小;I为未进行攻击的原始干净图像;MAX_I是指图像采样点颜色的最大值,一般为255。

3.2 参数对比实验

本节对总损失函数中的2个参数值α、β进行对比实验,简要分析2个不同参数取值对模型攻击效果的影响。根据特征损失和扰动损失取值大小[25],本文选用α=[0.000 1,0.000 2]和β=0.05作为初始值,然后在此区间范围内上下浮动取值进行实验,设置6组参数进行测试,分别计算对抗样本的峰值信噪比和均方误差,判断样本的图像质量,具体计算结果如表1所示。当PSNR值越大时,MSE值越小,图片质量越好。

表1 对抗样本的图像质量指标

Table 1 Image quality indicators of adversarial samples

参数组合	PSNR/dB	MSE
α= [0.000 1,0.000 2], β=0.005	7.2	2 035
α=[0.000 1,0.000 2], β=0.05	20.2	423
α=[0.000 1,0.000 2], β=0.5	27.3	110
α=[0.01,0.02], β=0.05	8.5	1 945
α=[0.001,0.002], β=0.05	12.5	1 456
α=[0.000 01,0.000 02], β=0.05	29.6	100

从表1可以看出,当α=[0.000 1,0.000 2],β=0.005和α=[0.01,0.02],β=0.05时,图像质量表现最差,其原因为当α取值过大或β取值过小时,导致特征损失增大而扰动损失减小,模型的扰动约束越小则图像中添加的扰动就越明显,扰动掩蔽性越差,图像质量越低。为进一步研究参数取值对攻击成功率的影响,根据表1结果选择图片质量较高的参数组合(PSNR>20)进行实验,具体实验结果如表2所示。

表2 不同参数组合的攻击成功率

Table 2 Attack success rate among different parameter combinations %

网络	ASR		
	α=[0.000 1, 0.000 2], β=0.05	α=[0.000 1, 0.000 2], β=0.5	α=[0.000 01, 0.000 02], β=0.05
ResNet 50	54.2	97.0	96.8
ResNet 101	70.2	94.6	83.3
ResNet 152	70.6	95.1	77.2
SEResNet 50	65.2	90.0	68.5
SEResNet 101	80.1	92.2	79.4
Attention 56	71.3	92.3	70.9
MobileNet	63.2	94.4	77.4

从表2可以看出,当参数β取值一定时,减小α值使得模型攻击效果更好。另外,当α值一定时,β取值越大攻击效果越好,表明在总损失中适当增加特征损失权重,减少扰动损失权重对模型效果有一定的提升。

由上述实验分析可得,在选取的6种参数设置中,当α=[0.000 01,0.000 02],β=0.05时,对抗样本的PSNR取得最大值,表示对抗样本的图像质量在此时取得最好的效果。在后续实验中,均设置α=[0.000 01,0.000 02],β=0.5。

3.3 单模型攻击实验

本节将G-MASK攻击方法与3种基线攻击方法进行单模型攻击对比实验,在实验过程中,首先利用多任务卷积神经网络(MTCNN)[26]对原始图像进行人脸检测,再将图像对齐并裁剪为112×112像素大小,图像的像素点总数为12 544。本文利用Grad-CAM算法选择输出热力图中权重值最大的5 000个像素点,对其进行攻击。其中,最大扰动ε=20,衰减系数μ=0.04,迭代次数T=40。利用3种基线攻击方法和本文提出的G-MASK方法在7种模型上实现攻击,并计算其攻击成功率,具体实验结果如表3所示,加粗表示最优数据,*表示白盒攻击。

从表3可以看出,G-MASK攻击成功率与3种基线方法相比得到显著提高,在白盒条件下,G-MASK方法的攻击成功率在各模型上均能达到98.7%以上。在黑盒条件下,G-MASK方法攻击成功率得到显著提升,在ResNet 50模型上生成的对抗样本用于SEResNet 101模型攻击时,G-MASK攻击成功率比基线方法MI-FGSM增长13.4个百分点,G-MASK方法与MI-FGSM相比在7个模型上的攻击成功率分别增长了12.4、13.0、7.0、12.9、19.4、14.7、6.6个百分点。根据实验结果可知,本文方法不仅能够在一定程度上提高白盒攻击成功率,而且还有效提升了黑盒攻击成功率,都优于现有的方法,充分表明本文所提方法具有一定的有效性。

表 3　在单模型攻击设置下不同方法的攻击成功率

Table 3　The success rate of attacks among different methods in single-model attack settings %

模型	攻击方法	攻击模型							平均值
		ResNet 50	ResNet 101	ResNet 152	SEResNet 50	SEResNet 101	Attention 56	MobileNet	
ResNet 50	FGSM	**90.1***	21.1	12.9	12.8	22.1	37.2	46.5	34.7
	I-FGSM	**94.6***	43.3	27.2	25.0	40.9	54.8	50.3	48.0
	MI-FGSM	**98.5***	60.8	44.5	48.8	56.8	62.7	52.6	60.7
	G-MASK	**99.4***	**77.4**	**63.2**	**78.1**	**70.2**	**63.1**	**60.3**	**73.1**
ResNet 101	FGSM	14.5	**92.8***	14.9	15.8	22.2	48.0	46.8	36.4
	I-FGSM	40.8	**95.2***	34.6	33.9	47.9	69.9	44.1	52.3
	MI-FGSM	48.5	**99.4***	41.8	40.5	52.6	74.9	46.2	57.7
	G-MASK	**73.6**	**100.0***	**76.0**	**75.2**	**66.8**	**75.9**	**47.4**	**70.7**
ResNet 152	FGSM	16.4	19.9	**93.5***	15.2	22.8	39.9	49.4	36.7
	I-FGSM	40.4	47.5	**94.3***	33.4	49.2	70.8	45.2	54.4
	MI-FGSM	45.2	52.7	**99.5***	40.1	55.5	72.6	53.4	67.4
	G-MASK	**76.4**	**78.8**	**99.7***	**78.1**	**69.3**	**72.8**	**56.4**	**74.4**
SEResNet 50	FGSM	15.4	24.7	16.4	**91.4***	22.5	37.2	49.9	36.7
	I-FGSM	40.3	51.4	34.8	**96.8***	51.5	48.8	56.5	54.3
	MI-FGSM	46.8	54.9	40.5	**98.7***	59.0	52.3	54.4	58.7
	G-MASK	**73.1**	**74.0**	**75.6**	**98.4***	**66.0**	**58.3**	**58.8**	**71.6**
SEResNet 101	FGSM	17.3	25.3	16.2	14.0	**92.6***	34.8	42.4	34.7
	I-FGSM	39.6	42.0	37.4	35.7	**97.3***	66.4	50.1	52.6
	MI-FGSM	46.2	50.9	42.6	39.8	**98.7***	67.4	62.5	58.7
	G-MASK	**79.7**	**76.0**	**77.7**	**67.2**	**100.0***	**68.6**	**67.6**	**78.1**
Attention 56	FGSM	5.0	10.0	6.2	4.9	8.2	**91.0***	39.9	23.6
	I-FGSM	25.2	33.6	21.7	19.0	32.4	**97.8***	55.4	40.7
	MI-FGSM	40.6	53.7	46.8	31.2	50.2	**99.7***	59.2	54.5
	G-MASK	**64.0**	**61.6**	**63.9**	**64.5**	**65.8**	**98.9***	**64.6**	**69.2**
MobileNet	FGSM	15.0	27.9	19.3	16.3	25.8	36.0	**91.4***	33.1
	I-FGSM	42.3	53.7	45.6	35.7	61.4	53.0	**97.2***	55.6
	MI-FGSM	46.7	58.4	52.4	40.2	64.2	58.5	**99.1***	60.0
	G-MASK	**64.0**	**54.4**	**58.3**	**63.0**	**66.4**	**66.3**	**99.8***	**66.6**

3.4　扰动集成模型攻击实验

本文虽然将上述方法用于单模型对抗攻击中能实现较优的黑盒攻击性能，但是在攻击有一定防御能力的模型时表现不佳。因此，本文在单模型攻击方法的基础上提出扰动集成模型攻击方法。将 ResNet 50、ResNet 101、ResNet 152、SEResNet 50、SEResNet 101、Attention 56、MobileNet模型每6个集成后进行白盒攻击，再利用生成的对抗样本对另1个模型进行黑盒攻击。本文分别使用3种基线对抗算法和G-MASK扰动集成算法进行攻击，实验结果如表4所示。其中，-ens代表扰动集成，*表示白盒攻击。

从表4可以看出，当进行扰动攻击集成后，4种算法在白盒攻击中均取得较好效果，成功率保持在98.5%以上，证明扰动集成能够提升白盒攻击成功率。在黑盒攻击中，G-MASK方法相比3种基线攻击方法攻击成功率始终提高10~20个百分点。因此，G-MASK方法的黑盒攻击效果要优于基线对抗攻击算法，具有更强的迁移性。

与表3中单一模型攻击成功率的实验数据相比，经过扰动集成后的G-MASK方法实现的黑盒攻击成功率平均提升13个百分点，充分证明扰动集成方法能够有效提升黑盒模型迁移能力和黑盒攻击成功率。

表4 在集成模型攻击设置下不同方法的攻击成功率

Table 4 The success rate of attacks among different methods in integrated model attack settings %

未参与集成的模型	SAR							
	攻击方法	ResNet 50	ResNet 101	ResNet 152	SEResNet 50	SEResNet 101	Attention 56	MobileNet
ResNet 50	FGSM-ens	50.7	98.5*	98.7*	98.6*	98.8*	98.5*	98.7*
	I-FGSM-ens	55.3	99.0*	99.0*	98.7*	99.3*	98.7*	98.9*
	MI-FGSM-ens	57.6	99.7*	99.8*	99.3*	99.9*	99.4*	99.0*
	G-MASK-ens	**60.8**	**100.0***	**100.0***	**99.9***	**100.0***	**100.0***	**99.9***
ResNet 101	FGSM-ens	98.7*	53.2	98.5*	98.8*	98.0*	98.8*	98.7*
	I-FGSM-ens	99.9*	55.6	98.7*	98.9*	98.4*	99.4*	99.2*
	MI-FGSM-ens	99.8*	59.1	98.8*	99.4*	99.5*	99.9*	99.8*
	G-MASK-ens	**100.0***	**69.2**	**99.7***	**100.0***	**99.9***	**100.0***	**99.9***
ResNet 152	FGSM-ens	98.8*	98.5*	56.8	99.7*	98.8*	98.5*	98.6*
	I-FGSM-ens	98.9*	98.9*	60.2	99.9*	99.3*	98.9*	98.3*
	MI-FGSM-ens	99.3*	99.2*	70.8	100.0*	100.0*	99.4*	99.4*
	G-MASK-ens	**99.8***	**100.0***	**73.3**	**100.0***	**100.0***	**100.0***	**100.9***
SEResNet 50	FGSM-ens	98.7*	98.9*	98.7*	59.2	98.9*	99.0*	98.8*
	I-FGSM-ens	98.8*	99.3*	99.1*	64.7	98.7*	99.4*	99.2*
	MI-FGSM-ens	99.5*	99.5*	99.9*	70.2	99.1*	99.6*	99.8*
	G-MASK-ens	**99.8***	**99.9***	**100.0***	**72.9**	**100.0***	**99.9***	**100.0***
SEResNet 101	FGSM-ens	99.1*	98.6*	98.5*	99.2*	62.1	98.6*	98.7*
	I-FGSM-ens	99.6*	99.2*	98.9*	99.4*	72.5	98.9*	99.4*
	MI-FGSM-ens	99.7*	99.8*	99.2*	99.9*	73.6	99.0*	99.7*
	G-MASK-ens	**100.0***	**100.0***	**99.9***	**99.9***	**74.2**	**99.8***	**100.0***
Attention 56	FGSM-ens	98.5*	99.2*	98.9*	98.8*	98.6*	63.8	99.3*
	I-FGSM-ens	99.4*	99.6*	99.3*	98.4*	98.9*	65.3	99.5*
	MI-FGSM-ens	99.8*	99.9*	99.9*	99.5*	99.6*	67.7	99.9*
	G-MASK-ens	**99.9***	**100.0***	**99.9***	**99.9***	**100.0***	**75.9**	**99.9***
MobileNet	FGSM-ens	98.8*	99.0*	99.0*	99.8*	98.7*	98.8*	64.8
	I-FGSM-ens	99.0*	99.8*	99.0*	99.9*	98.9*	99.3*	70.6
	MI-FGSM-ens	99.4*	100.0*	99.4*	99.9*	99.2*	99.6*	72.1
	G-MASK-ens	**100.0***	**100.0***	**100.0***	**100.0***	**100.0***	**100.0***	**75.6**

3.5 高斯滤波的有效性

高斯滤波的加入是为了使模型集成过程中减少不同模型干扰噪声之间的差异,同时提升对抗样本运用到其他黑盒模型上的攻击成功率,增强对抗样本掩蔽性。扰动集成模型使用高斯滤波和未使用高斯滤波生成的对抗样本准确度、Cosine相似度、L2距离如图3所示。从图3可以看出,添加高斯滤波后G-MASK生成的对抗样本攻击成功率在40次迭代时达到稳定,比未加入高斯滤波时MASK在20次迭代达到稳定耗费的时间更多,但是它们相差结果不大且最终都能在40次迭代时具有相同的攻击效果。此外,加入高斯滤波后生成的对抗样本余弦相似度显著增加但L2距离无明显变化,表明生成的对抗样本能够在保持相同攻击成功率的条件下降低与输入样本的图像差异,提高图像质量,在一定程度上增加对抗样本的掩蔽性,同时增大在模型防御场景下的攻击成功率。

另外,本文对未添加高斯滤波攻击方法(MASK)和添加高斯滤波攻击方法(G-MASK)以及3种基线攻击方法生成的对抗样本进行对比分析,生成的对抗样本如图4所示。从图4可以看出,在算法中添加高斯滤波之后生成的对抗样本比原算法生成的对抗样本掩蔽性更强,人眼几乎不可察,证明高斯滤波的有效性。G-MASK攻击方法生成的对抗样本掩蔽性明显优于其他基线攻击方法,图像质量最高。本文方法所生成的对抗样本能够在保证较高峰值信

噪比的同时，即图片具有较少失真的情况下达到较高的攻击成功率，证明了本文所提方法的有效性。

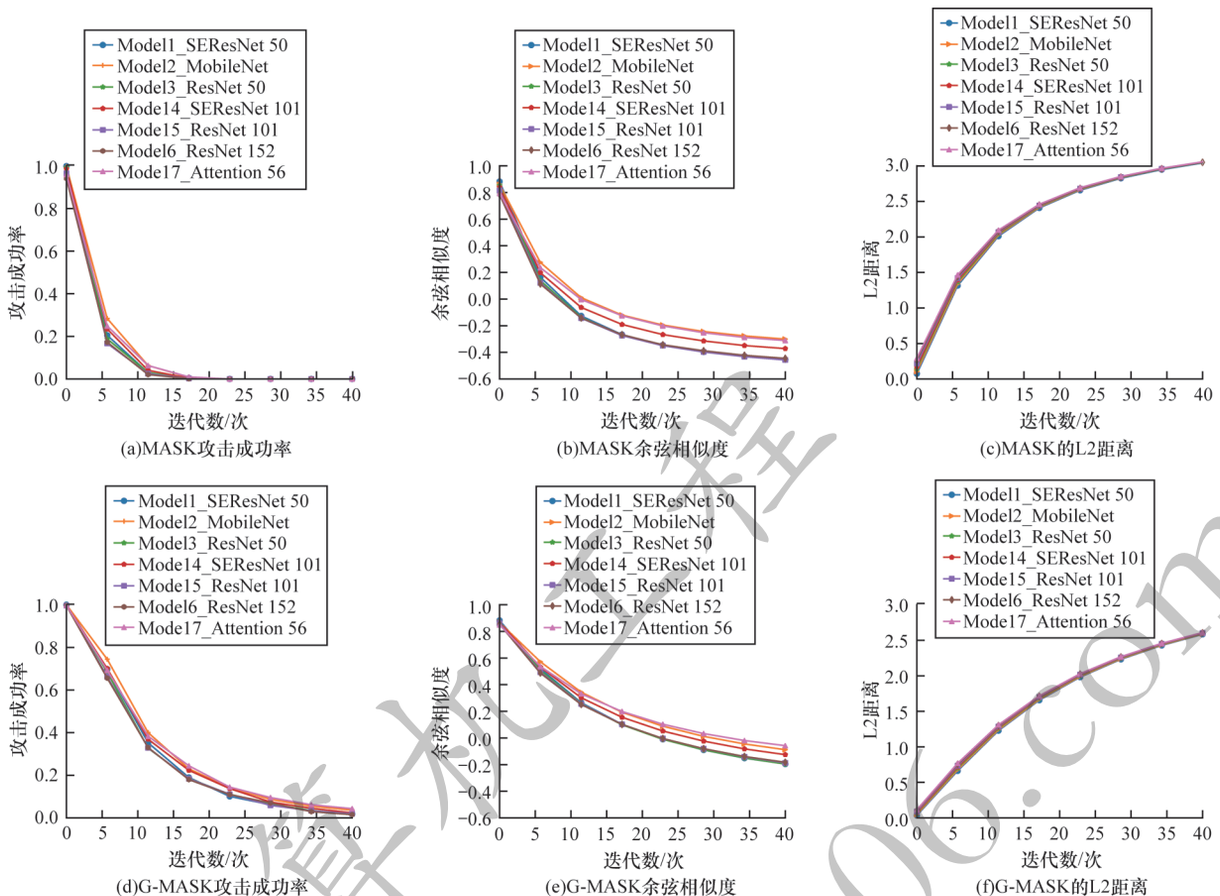

图 3　MASK 和 G-MASK 的攻击成功率、余弦相似度、L2 距离对比

Fig.3　Comparison of MASK and G-MASK of accuracy, cos-similarity, L2 distance

图 4　不同攻击方法生成的对抗样本

Fig.4　Adversarial samples generated by different attack methods

4　结束语

本文提出一种结合高斯滤波与人脸 MASK 的人脸攻击方法。通过引入 Grad-CAM 方法选择 MASK 区域进行扰动添加，限制扰动添加区域；此外，使用扰动集成模型提高黑盒攻击迁移能力，增强黑盒模型攻击成功率；同时对生成的扰动进行高斯平滑处理，增强扰动掩蔽性，使用最小且掩蔽性最强的扰动来达到最高的黑盒攻击成功率。实验结果表明，与其他攻击方法相比，本文方法具有更高的攻击成功

率和更优的掩蔽性,黑盒表现更为优秀。下一步将对算法速度的提高方法以及物理世界中更复杂的多人脸识别模型进行研究,在保持较高攻击成功率的前提下加快对抗样本生成速度。

参考文献

[1] LI H F, HUANG H K, CHEN L, et al. Adversarial examples for CNN-based SAR image classification: an experience study[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2021, 14: 1333-1347.

[2] 陈晓楠,胡建敏,张本俊,等. 基于模型间迁移性的黑盒对抗攻击起点提升方法[J]. 计算机工程,2021,47(8):162-169.
CHEN X N, HU J M, ZHANG B J, et al. Black box adversarial attack starting point promotion method based on mobility between models[J]. Computer Engineering, 2021, 47(8): 162-169. (in Chinese)

[3] 柴梦婷,朱远平. 生成式对抗网络研究与应用进展[J]. 计算机工程,2019,45(9):222-234.
CHAI M T, ZHU Y P. Research and application progress of generative adversarial networks [J]. Computer Engineering, 2019, 45(9): 222-234. (in Chinese)

[4] SHEN M, YU H, ZHU L H, et al. Effective and robust physical-world attacks on deep learning face recognition systems[J]. IEEE Transactions on Information Forensics and Security, 2021, 16: 4063-4077.

[5] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84-90.

[6] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[EB/OL]. [2022-11-05]. https://arxiv.org/pdf/1409.1556.pdf.

[7] SZEGEDY C, VANHOUCKE V, IOFFE S, et al. Rethinking the inception architecture for computer vision[C]//Proceedings of Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2016: 1-10.

[8] SZEGEDY C, IOFFE S, VANHOUCKE V, et al. Inception-v4, Inception-ResNet and the impact of residual connections on learning[C]//Proceedings of the 31st AAAI Conference on Artificial Intelligence. [S. l.]: AAAI Press, 2017: 4278-4284.

[9] SHARIF M, BHAGAVATULA S, BAUER L, et al. Accessorize to a crime: real and stealthy attacks on state-of-the-art face recognition[C]//Proceedings of ACM SIGSAC Conference on Computer and Communications Security. New York, USA: ACM Press, 2016: 1528-1540.

[10] 姜妍,张立国. 面向深度学习模型的对抗攻击与防御方法综述[J]. 计算机工程,2021,47(1):1-11.
JIANG Y, ZHANG L G. Survey of adversarial attacks and defense methods for deep learning model[J]. Computer Engineering, 2021, 47(1): 1-11. (in Chinese)

[11] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples[EB/OL]. [2022-11-05]. https://arxiv.org/pdf/1412.6572.pdf.

[12] KURAKIN A, GOODFELLOW I J, BENGIO S, et al. Adversarial machine learning at scale[EB/OL]. [2022-11-05]. https://arxiv.org/pdf/1611.01236.pdf.

[13] DONG Y P, LIAO F Z, PANG T Y, et al. Boosting adversarial attacks with momentum[C]//Proceedings of Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2018: 9185-9193.

[14] ZHANG Y K, JIANG Z Y, VILLALBA J, et al. Black-box attacks on spoofing countermeasures using transferability of adversarial examples[C]//Proceedings of Conference on the International Speech Communication Association. Washington D. C., USA: IEEE Press, 2020: 4238-4242.

[15] TAIGMAN Y, YANG M, RANZATO M, et al. DeepFace: closing the gap to human-level performance in face verification[C]//Proceedings of Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2014: 1-10.

[16] SUN Y, WANG X G, TANG X O. Deep learning face representation from predicting 10, 000 classes [C]//Proceedings of Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2014: 1-10.

[17] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition [C]//Proceedings of Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2016: 770-778.

[18] HU J E, SHEN L, SUN G. Squeeze-and-excitation networks[C]//Proceedings of Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2018: 7132-7141.

[19] WANG F, JIANG M Q, QIAN C, et al. Residual attention network for image classification [C]//Proceedings of Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2017: 1-10.

[20] WANG P S, CHENG J. Accelerating convolutional neural networks for mobile applications [C]//Proceedings of Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2017: 1-10.

[21] SELVARAJU R R, COGSWELL M, DAS A, et al. Grad-CAM: why did you say that? Visual explanations from deep networks via gradient-based localization[EB/OL]. [2022-11-05]. https://arxiv.org/pdf/1610.02391.pdf.

[22] MOOSAVI-DEZFOOLI S M, FAWZI A, FAWZI O, et al. Universal adversarial perturbations [C]//Proceedings of Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2017: 1765-1773.

[23] HUANG G B, MATTAR M A, BERG T L, et al. Labeled faces in the wild: a database for studying face recognition in unconstrained environments[EB/OL]. [2022-11-02]. http://download.xuebalib.com/1jeugxZEdrG3.pdf.

[24] KINGMA D P, BA J. Adam: a method for stochastic optimization[EB/OL]. [2022-11-05]. https://arxiv.org/abs/1412.6980.

[25] WEI X X, LIANG S Y, CHEN N, et al. Transferable adversarial attacks for image and video object detection[C]//Proceedings of the 28th International Joint Conference on Artificial Intelligence. New York, USA: ACM Press, 2019: 954-960.

[26] ZHANG K P, ZHANG Z P, LI Z F, et al. Joint face detection and alignment using multitask cascaded convolutional networks[J]. IEEE Signal Processing Letters, 2016, 23(10): 1499-1503.

编辑 薛晋栋