计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.241100104

基于消除语义特征的图像篡改定位模型对抗攻击

蒋伟豪　刘　波
重庆邮电大学图像认知重庆市重点实验室　重庆 400065
(2021211795@stu.cqupt.edu.cn)

摘　要　目前,公众对于日新月异的图像篡改技术越来越担忧,因为它会引发伦理和安全问题。利用深度神经网络可以定位图像篡改区域。然而,随着深度神经网络的发展,针对它的对抗性攻击也层出不穷,这些攻击方法也促进了模型的鲁棒性研究。现有的对抗攻击方法主要关注篡改痕迹特征,然而不同图像篡改定位模型关注的篡改痕迹特征有所不同,导致对抗攻击的迁移能力不足。由于卷积神经网络或 Transformer 网络也能够提取语义特征,而图像篡改定位模型往往将这些模型作为基线模型,因此模型在提取篡改特征时会不可避免地提取到部分语义特征。为了提高对抗样本的泛化能力,提出一种攻击方法,重点关注消除篡改图像的语义特征,训练一个语义分割网络作为攻击目标,提出一种攻击中间语义特征的损失函数,使得模型难以识别出图像篡改部分的语义信息。这种攻击方法具有较高的迁移能力,可以更好地隐藏扰动并生成更具攻击性的对抗样本,在多种实验下被证明可以攻击绝大多数现有模型并优于其他对抗攻击方法,并为图像篡改定位任务提供了更新颖的见解。

关键词:对抗攻击;深度学习;图像篡改定位
中图分类号　TP309;TP391;TP183

Attacking Image Manipulation Localization Model by Eliminating Semantic Features

JIANG Weihao and LIU Bo
Chongqing Key Laboratory of Image Cognition,Chongqing University of Posts and Telecommunications,Chongqing 400065,China

Abstract　At present,the public is increasingly concerned about the image tampering technology because it will cause ethical and security issues. Deep neural networks can be used to locateimage tampering areas. However,with the development of deep neural networks,adversarial attacks against them have also developed,and these attack methods have also promoted the research on the robustness of the model. Existing adversarial attack methods mainly focus on tampering trace features,but different Image Manipulation Localization models focus on different tampering trace features,resulting in insufficient migration ability of adversarial attacks. Since convolutional neural networks or Transformer networks can also extract semantic features,and Image Manipulation Localization models often use these models as baseline models,which would inevitably extract some semantic features when extracting tampering features. In order to improve the generalization ability of adversarial samples,a attack method is proposed,focusing on eliminating the semantic features of tampered images,training a semantic segmentation network as the attack target,and proposing a loss function for attacking intermediate semantic features,making it difficult for the model to identify the semantic information of the tampered part of the image. This attack method has better transfer ability,can hide perturbations and generate more aggressive adversarial samples. It has been proven in multiple experiments that it can attack most existing models and outperform other adversarial attack methods,and provides novel insights for the image manipulation localization.

Keywords　Adversarial attack,Deep network,Image manipulation localization

1　引言

随着图片编辑软件的普及,图片篡改已经成为一种简单且低成本的行为。一些恶意篡改图片在互联网上传播,会对网络信息安全造成危害。应用图像篡改定位模型能够实现自动识别篡改,目前已经有很多深度神经网络可以有效地检测篡改图片,并分割出图片中被篡改的部分,如 RRU-Net[1],PSCC-Net[2],MVSS-Net[3]等。深度神经网络虽然有效,但也存在容易受到攻击的弱点。图 1 展示了一张被篡改的图片,其被输入到图像篡改定位模型后,模型可以准确地分割出篡改部分,而如果使用特殊的算法在图片上添加一些被设计过的微小扰动[4],深度神经网络就难以像往常一样正确地分割出篡改部分。

这种为图片添加扰动以攻击模型的方法被称为对抗攻击[4],添加扰动后的图片称为对抗样本[4]。随着深度神经网络的发展,针对它的对抗性攻击也层出不穷,这些攻击方法也能够促进模型的鲁棒性研究[5]。对抗样本的攻击性能和泛化性能可以评估攻击方法是否有效。对于图像篡改定位任务来说,现有的对抗攻击方法[6-8]总是将图像篡改定位模型当做一般的分割模型进行研究和攻击,而忽视了图像篡改定位中一个

基金项目:重庆市自然科学基金面上项目(CSTB2023NSCQ-MSX0341)
This work was supported by the Natural Science Foundation of Chongqing(CSTB2023NSCQ-MSX0341).
通信作者:刘波(boliu@cqupt.edu.cn)

重要的问题，即区分篡改图像中的语义特征和篡改痕迹特征。

对于图像篡改定位模型，不同的模型学习到的篡改痕迹也不尽相同。只根据某个模型的篡改痕迹特征进行对抗攻击，对抗样本的泛化能力也较弱。虽然模型应该学到尽可能多的篡改痕迹特征而非只专注篡改部分的语义特征，但由于一些常用的神经网络结构如卷积神经网络[9]或者 Transformer[10]等，总是本能地关注语义信息，且图像篡改定位模型常常基于这些基线网络构建，因此模型在训练过程中会不可避免地学习到部分语义特征，这是图像篡改定位模型的共同弱点。

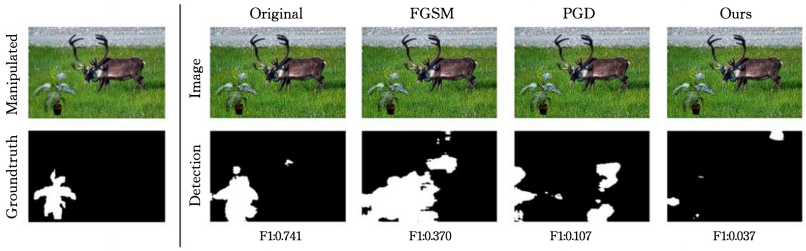


图 1 针对图像篡改定位模型的对抗攻击示意图

Fig. 1 Schematic diagram of adversarial attack on image tampered localization model

为了让对抗攻击有足够的泛化能力，本文提出了一种基于迁移攻击的消除语义对抗攻击方法。准备一个语义分割模型作为代理模型，然后通过攻击这个白盒模型获得对抗样本，这些对抗样本可以用来攻击其他模型。这种方法要求生成扰动的算法足够好，因此本文也提出了一种用于清除语义特征的对抗攻击损失函数。这种对抗攻击方法在和其他同类型方法比较下均体现出了较好的攻击性能和迁移性能。本文的主要贡献包括：

1）为对图像篡改定位任务的对抗攻击提供一个优秀的代理模型，它是一个语义分割模型，并保留了注意力机制，能够很好地提取出篡改图像篡改部分的语义特征。

2）为该代理模型的语义消除对抗攻击设计了一种新的损失函数。该损失函数的算法基于消除模型注意力，使用该损失函数生成对抗样本可以更好地消除指定的语义特征。

2 相关工作

2.1 图像篡改定位模型

图像篡改定位是一项分割任务，目的是在像素级别定位篡改区域。RRU-Net[1]提出实现一个循环残差模块来定位篡改痕迹。PSCC-Net[2]设计了一个新颖的 SCCM 模块来捕获空间和通道相关性，以实现更好的泛化性能。MVSS-Net[3]使用多视图特征学习来获得泛化特征。除了这些模型之外，还有许多优秀的图形篡改定位模型，如 NEDB-Net[11]，IML-ViT[12]，MMFusion[13]，HiFi-Net[14]等，每个模型都有自己的特点。然而，上述模型仍然是基于卷积神经网络[9]或 Transformer[10]的结构，这些结构也被广泛应用于语义特征识别的任务，证明了这种网络结构在训练的过程中也会不可避免地提取出图像的语义特征，这些特征可以被看作是这些模型共同的弱点。

2.2 对图像篡改定位模型的对抗攻击

对抗攻击可以通过添加微小扰动生成令深度神经网络误判的对抗样本，以此来揭示深度神经网络的弱点。对抗样本的干扰对人类来说往往是难以察觉的，但深度神经网络会以很高的置信度错误地检测到它们。当前的对抗攻击算法可以分为两类，即白盒攻击和黑盒攻击。

白盒攻击方法即根据模型内参数来生成对抗样本。白盒攻击的应用背景是被攻击的模型的所有参数和结构都被攻击者知晓，因此攻击者可以根据模型的信息来生成有针对性的对抗样本。FGSM[4]是一种传统的一步攻击方法，它使用输入图像的损失函数的逆梯度来生成干扰。PGD[15-16]是一种迭代攻击，它以较小的步长执行梯度更新，并将更新后的对抗样本剪辑到有效范围内。其他方法如 CW[17]，JSMA[18]，MI-FGSM[19]也被广泛使用。

当受害者模型的信息已知且其访问权限不受限制时，上述攻击成功率很高，而现实场景中，攻击者很难知晓模型内部的结构和参数，也很难高频率不受限制地访问模型。因此，攻击者也需要黑盒攻击[20]。黑盒攻击即将被攻击的模型当成一个不能无限制访问也不知晓内部结构的黑盒模型[20]。

黑盒攻击的一种主流实现方法为迁移攻击[21]，即攻击者在指定的白盒模型中执行白盒攻击，生成的对抗样本对其他模型也具有攻击性，此时被指定的白盒模型被称为代理模型[21]。为了让对抗样本具备足够的可迁移性，除了扰动生成算法本身需要足够优秀之外，代理模型的选取也至关重要。为了增强对抗样本的可迁移性，目前的黑盒攻击方法扩展了许多扰动生成算法，如基于梯度优化的 AoA 算法[22]在图像分类中试图消除图像在模型下的注意力，T-SEA 算法[23]在物体识别中利用自集成策略增强对抗样本的泛化能力。这些算法虽然取得了相对有效的效果，但并未注重对代理模型的选取。在图像篡改定位任务领域也已有一些对抗攻击方法，如针对图像篡改定位的深度对抗攻击[6]、针对基于卷积网络的对图像篡改定位对抗攻击方法[7]，以及文献[8]中提出的 Frequency-aware GAN。这些方法都基于黑盒攻击环境提出了自己的见解和方法，但都没有注意到图像篡改定位任务和一般语义分割任务的差别，以及模型中语义特征和图像篡改特征的差别，它们把图像篡改定位模型当做一般的分割模型来讨论攻击方法。

3 消除语义对抗攻击

为了更好地攻击图像篡改定位模型，本文提出了一种基于迁移的语义消除对抗攻击方法。如图 2 所示，这种方法可以分为两个步骤，提取语义特征和消除语义特征。在提取语义特征过程中，一种语义分割模型将被建立作为代理模型，篡改图片通过语义分割模型后，模型将提取图片的语义特征。在消除语义特征过程中，图片将根据本文提出的损失函数生成扰动，这种扰动可以消除图片的语义特征，进而让其他的图像篡改定位模型无法识别图片的篡改部分。

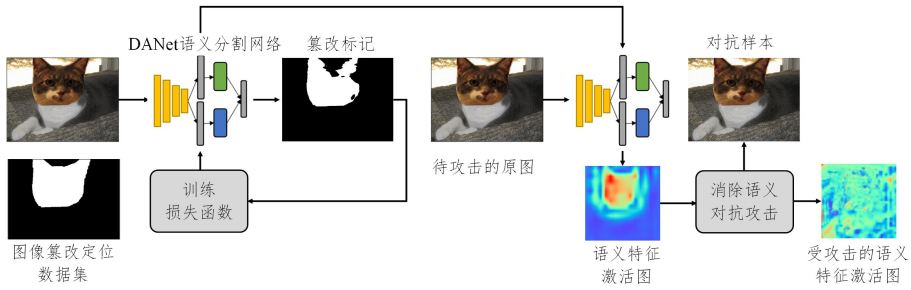


图 2　消除语义特征的对抗攻击方法示意图

Fig. 2　Schematic diagram of adversarial attack by eliminating semantic features

3.1　提取语义特征

由于卷积神经网络[9]和 Transformer[10]常用于语义分割的方法，其已经被证明了能够提取语义特征，而大多数图像篡改定位模型也都是从这两种模型中选择基线模型，因此，图像篡改定位模型在训练过程中会不可避免地学习到被篡改部分的语义特征。

因此，如果训练一个专门用于语义分割的模型，利用图像篡改定位数据集作为语义分割数据集来训练模型，并使用这个“过拟合”的模型作为代理模型来生成对抗样本，那么这个模型也能够提取出图像篡改部分的语义特征。

对于语义分割模型的选择，模型的结构应该类似于图像篡改定位模型的结构。对于大多数图像篡改定位模型，往往会考虑注意力机制和 Unet 结构，例如 MVSS-Net[3]，PSCC-Net[2]，ARNet[24]，SPAN[25]等。因此，本文选择采用 DA-Net[26]作为代理模型。DANet[26]是一个经典的语义分割模型，它以 ResNet[27]为基线模型，并在模型末尾添加了双重注意力机制的模块，这使模型结构简单但能够有效提取图像的语义特征。

3.2　消除语义特征

在提出了一个语义分割模型作为代理模型之后，需要设计一种合适的方法来攻击这个特殊的模型，以消除图像的语义特征。传统的对抗攻击方法总是直接使用模型输出作为欺骗的目标，而模型中间层的特征图包含的信息更多，欺骗特征图可以使扰动抹去更多的图像语义特征，从而增强对抗样本的泛化性。图 3 展示了不同训练模型的输出，RRU-Net，PSCC-Net 和 MVSS-Net 都是图像篡改定位模型，但 DANet 是语义分割模型。可以看出，对于相同的输入，不同训练模型在同一任务上的输出总是相似的，但特征图却有很大的不同，其明显拥有更多的信息，从特征图中寻找模型弱点更为可靠。

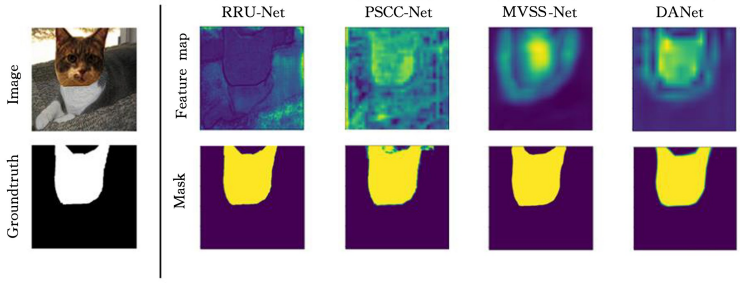


图 3　不同图像篡改定位模型的中间层特征图

Fig. 3　Intermediate layer features of various image tampered localization model

对抗攻击可以表示为：

$$\begin{aligned}&\text{find } \Delta x\\&\text{s. t. } f(x)\neq f(x+\Delta x)\\&\quad \|\Delta x\|<\varepsilon\\&\quad x_{\text{adv}}=x+\Delta x\end{aligned}\tag{1}$$

其中，x 是模型的输入，Δx 是扰动。Δx 受到 $\|\cdot\|$ 的限制。对于大多数现有的对抗攻击方法，攻击的目标往往是 $f(x)$。只要 $f(x_{\text{adv}})$的值不同于 $f(x)$，攻击自然有效。诚然这是一种直接而直观的方法，但 $f(x)$所蕴含的关于模型的信息是有限的。对于大多数深度神经网络来说，最终的 $f(x)$都会被简单的卷积层或者全连接层降维，这会丢失大量的特征信息。为了实现 $f(x)\neq f(x_{\text{adv}})$，对于有多个中间层的深度神经网络，只需要对中间层的特征图也进行攻击，就能达到最终 $f(x_{\text{adv}})$被欺骗的效果。对于具有多个中间层的深度神经网络，可以将其表示为：

$$f(x)=f_2(f_1(x))\tag{2}$$

此处按层划分将深度神经网络 $f(x)$分为 f_1 和 f_2 两个部分，其中 f_1 是 $f(x)$在某中间层之前的部分，f_2 是 $f(x)$在该中间层之后的部分。$f_1(x)$是中间层输出的特征图。将特征图直接输入到 f_2 中，可以得到模型 $f(x)$的最终输出。

如果攻击 f_1，那么攻击目标可以写成：

$$\text{s. t. } f_1(x)\neq f_1(x+\Delta x)\tag{3}$$

由于对抗样本已经在 f_1 部分使模型失效，因此可以忽略后续部分 f_2，因为 f_2 无法从因为被攻击而失去原有信息的 $f_1(x_{\text{adv}})$中提取出原本的正确特征。因此，攻击任务在此处从攻击 f 变为攻击 f_1。由于 $f_1(x)$具有更丰富的特征信息，因此需要一个新的损失函数来攻击它。

由于 f_1 输出的并不是一个简单的标签而是带有复杂信息

的特征图，而特征图的取值范围和分布都非常不均匀，直接比较$f_1(x)$和$f_1(x_{adv})$并不合理。

本文提出一种处理方式是将特征图的注意力分散。以降维后的特征图的方差作为损失函数，这样在迭代过程中，特征图的分布就会趋于均匀，失去焦点。为了获取更多的特征信息，选择从模型输出层上一层获取特征图。由于各个模型中特征图的大小不一样，特征图的通道也很难统一，所以需要对特征图进行简单的处理，先通过求和对特征图进行降维，压缩至单通道图，然后进行归一化，最后调整所有特征图为[256, 256]的大小。

为了欺骗受害者模型，对抗攻击中用于计算梯度的损失函数通常为模型训练的损失函数，而本文利用计算特征图的方差作为损失函数。设 $x=x_{ori}$，则攻击过程可以完整描述为：

$$
\begin{aligned}
M&=\frac{\sum f_1(x)}{N}\\
J(\theta,x)&=\frac{\sum(f_1(x)-M)^2}{N}\\
x'_{t+1}&=Clip_{x,\varepsilon}\left(x_t'+\varepsilon*\frac{\nabla_x J(\theta,x)}{\|\nabla_x J(\theta,x)\|_1/N}\right)
\end{aligned}
\tag{4}
$$

其中，x_t' 为迭代过程中的对抗样本，梯度通过其平均值 l_1-norm 进行归一化，即 $\|g(x)\|_1/N$，其中 N 是图像的大小。f_1是没有最后一层的模型，$f_1(x)$是倒数第二层的特征图。这些特征图的通道数总是不唯一的，应该先将它们降维压缩成一个通道。M代表特征图的平均期望，而 $J(\theta,x)$实际上为特征图的方差。如果$f_1(x)$的方差能变小，则模型的特征图将更模糊，这使得模型难以聚焦于正确特征。如果攻击中间层的特征图，那么将不需要关注该层之后的模型结构，因为无论什么样的模型结构都不可能从无内容的特征图中得到正确的特征。

图 4 展示了在语义特征消除的过程中图像篡改定位模型检测的精度的变化，可见随着语义特征被消除，图像篡改定位模型也难以识别篡改部分。由于这种方法可以消除图片的语义特征，因此生成的图片不仅可以攻击图像篡改定位模型，而且攻击的效果还和模型学习到的语义特征的多少有关，模型越是注重图片的语义特征，这种方法可以达到的攻击效果就越强。因此本文的方法对某个模型的攻击效果也在一定程度上反映了该模型是否真的学到了正确的篡改痕迹。

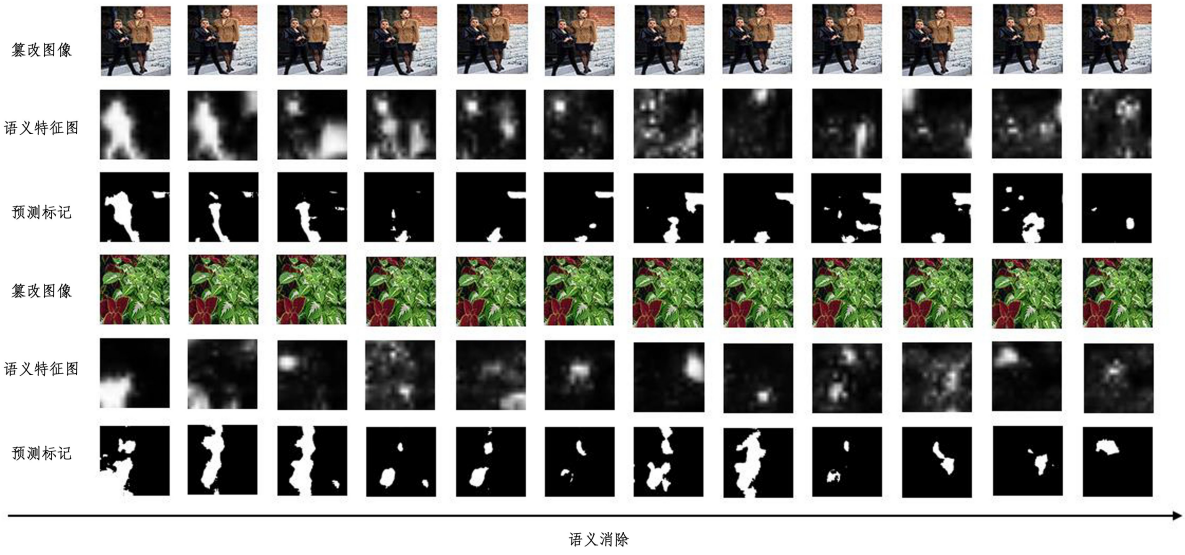


图 4 语义消除及中间层特征图模糊化的过程

Fig. 4 Process of semantic elimination and blurring of intermediate feature map

4 实验及结果分析

本章将评估本文选择的语义分割代理模型和损失函数在黑盒攻击中的表现。首先训练代理模型，然后在攻击过程中将图像输入代理模型，使用不同的方法生成对抗样本。在黑盒设置中，使用基于可迁移性的攻击来选择不同的模型作为代理模型并获得对抗样本，然后直接使用这些对抗样本攻击另一个未知模型。将语义消除对抗攻击方法在 CASIAv1 和 CASIAv2 数据集上进行评估，并将本文方法与现有的对抗攻击方法 FGSM，PGD，advGAN 和 Frequency-aware GAN 进行比较。最后，将展示本文方法生成的对抗样本对图像篡改定位模型的攻击性能与该模型本身的泛化能力之间的相关性，所提方法在 HiFi-Net[14]，MMFusion[13]，TruFor[28]，FOCAL[29]，NEDB-Net[11]，CAT-Net[30]，IML-ViT[12]，PSCC-Net[2]，MVSS-Net[3]，DANet[26]上进行实验。

4.1 实验数据库及评估方法

本实验选择在图像篡改数据集 CASIAv1 和 CASIAv2[6] 上评估损失函数。CASIAv1 包含 921 张来自 Corel 数据集[15]的带有拼接操作的图像。CASIAv2 是一个自然图像篡改数据集，包含 5 123 张带有复制移动和拼接操作的图像。所有数据集都提供了篡改区域的掩码标签。实验将用跨数据集测试来直接表示模型的泛化能力，跨数据集包括的图像篡改数据集有 CASIAv1，Columbia，Coverage 和 NIST16。

关于选择代理模型的对照实验，本文选择了 RRU-Net，PSCC-Net 和 MVSS-Net 这 3 个图像篡改定位模型。这些模型都是在 CASIAv2 数据集上训练的。将本文方法生成的对抗样本与这些模型生成的对抗样本进行比较，以证明对于图像篡改定位模型，语义消除对抗攻击方法已经具有足够的攻击性能。对于用于评估泛化性能的模型，选择了近几年新颖的图像篡改定位模型，包括 HiFi-Net[18]，MMFusion[29]，TruFor[28]，FOCAL[29]，NEDB-Net[11]，CAT-Net[30]，IML-ViT[12]，PSCC-Net[2]，NEDB-Net[11]，MVSS-Net[3]。实验还将在语义分割模型 DANet 上评估泛化性能，以进一步判断评估是否合理。

将数据集上模型输出与真实标签的平均 F1-score 作为模型的准确率指标。实验还使用 AR 来比较语义消除对抗攻击生成的对抗样本的攻击性能。AR 的计算式如式(7)所示：

$$AR(x_{ori},x_{adv})=\frac{F_1(f(x_{ori}),y)-F_1(f(x_{adv}),y)}{F_1(f(x_{ori}),y)} \tag{7}$$

其中，F_1 为 F1-score，x_{ori} 为原图，x_{adv} 为对抗样本，y 为标签，$f(x)$ 代表模型的输出。AR 可以用来评估模型被攻击的程度。

4.2 对抗攻击迁移攻击性能比较

本节进行对照实验来评估语义消除对抗攻击方法的可迁移性。使用 CASIAv2 数据集和 CASIAv1 数据集分别训练两个 DANet 作为代理模型，一个用于 CASIAv2 的攻击，另一个用于 CASIAv1 的攻击，并将这些代理模型与其他图像篡改定位模型进行比较。结果如表 1 所列。用于生成对抗样本的模型表示为代理模型，受害模型用于计算 F1-score 以评估攻击性能。所有对抗样本均由 3.3 节中提出的基于梯度的方法生成，并且所有对抗样本生成扰动都被限制为 $\varepsilon=0.05$。为了更好地进行比较，还添加了白盒攻击的结果。最后，比较代理模型在其他模型中的平均攻击率(AR)和平均转移攻击率(AR')，AR'忽略了白盒攻击的结果，只计算了迁移攻击的攻击率。从表 1 中可以观察到，无论是整体攻击能力，还是只考虑转移攻击的可转移性，DANet 都是最优的代理模型。

表 1 语义消除攻击的黑盒攻击实验

Table 1 Black-box attack experiment of semantic elimination attack

Surrogate model	Victimmodel	F_1		AR		Average AR		Average AR'	
		CASIAv1	CASIAv2	CASIAv1	CASIAv2	CASIAv1	CASIAv2	CASIAv1	CASIAv2
PSCC-Net[2]	PSCC-Net	0.1392	0.1283	70.49	84.65	36.98	45.71	20.22	26.22
	MVSS-Net	0.3719	0.6925	11.78	19.46				
	RRU-Net	0.2436	0.5972	28.67	32.99				
MVSS-Net[3]	PSCC-Net	0.2480	0.5498	47.44	34.27	48.04	44.20	41.91	31.25
	MVSS-Net	0.1673	0.2571	60.31	70.10				
	RRU-Net	0.2173	0.6395	36.37	28.24				
RRU-Net[1]	PSCC-Net	0.2233	0.3283	52.69	60.75	29.25	44.00	29.58	50.35
	MVSS-Net	0.3942	0.5164	6.47	39.94				
	RRU-Net	0.2439	0.6123	28.60	31.30				
DANet[25]	PSCC-Net	0.2088	0.4245	55.76	49.25	47.51	51.41	47.51	51.41
	MVSS-Net	0.2273	0.2689	46.07	68.73				
	RRU-Net	0.2026	0.5680	40.70	36.26				

由于语义消除对抗攻击方法只是消除了图片的语义特征而并不直接对篡改特征进行消除，因此不同模型在遇到此方法生成的对抗样本时，攻击率也与这些模型检测篡改痕迹的能力相关。如果消除语义对抗攻击在某个模型上的攻击率较低，那么不仅证明了该模型对扰动具有很好的鲁棒性，而且也证明了该模型真正学习到了图像的篡改痕迹信息，这样的模型才是在实际环境中可用的图像篡改定位模型。

此外，实验选取了几个现有的模型进行评估。一般认为，图像篡改定位模型学习到的篡改痕迹特征越多，对不同场景的适应性就越强，因此在跨数据集上的表现也应该更好。在本次实验中，以跨数据集测试下不同模型的平均 F1-score 作为评价模型泛化能力的标准。如表 2 所列，利用 CASIAv2 中经过语义消除的对抗样本对这些图像篡改定位模型进行攻击，以这些模型下对抗样本的攻击率作为评价模型学习到的篡改痕迹特征占比的标准。

表 2 不同图像篡改定位模型的泛化能力

Table 2 Generalization ability of different image tampering localization models

Model	CASIAv1	CASIAv2	Coverage	NIST16	Average F_1
FOCAL[27]	0.898	0.981	0.863	0.737	0.870
HiFi-Net[7]	0.616	0.912	0.801	0.850	0.795
MMFusion[6]	0.784	0.888	0.663	0.430	0.691
TruFor[26]	0.737	0.859	0.600	0.399	0.648
IML-ViT[5]	0.658	0.836	0.425	0.339	0.565
RRU-Net[1]	0.841	0.915	0.199	0.262	0.554
PSCC-Net[2]	0.363	0.935	0.498	0.357	0.538
NEDB-Net[4]	0.511	0.753	0.463	0.291	0.505
CAT-Net[28]	0.710	0.799	0.107	0.242	0.465
MVSS-Net[3]	0.452	0.638	0.453	0.292	0.459
DANet[25]	0.298	0.371	0.198	0.139	0.251

图 5 展示了不同模型的受攻击率。以模型下对抗样本的攻击率为横坐标，以模型的跨数据集平均 F1-score 为纵坐标，绘制出如图 6 所示的散点图，显然可以得出两者之间存在线性相关性。这不仅进一步证明了语义消除对抗攻击方法是优秀的，也暗示了图像篡改定位模型的性能和它关注语义特征的程度很可能有一定程度上的负相关性。

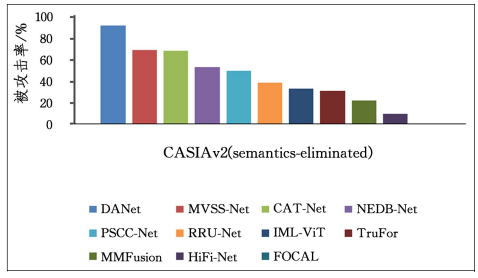


图 5 不同模型的被攻击率

Fig. 5 Various model's attack rate

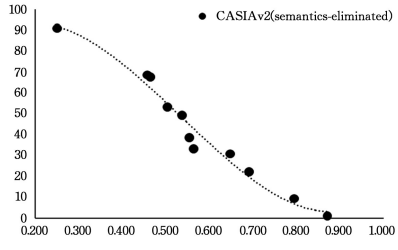


图 6 模型的被攻击率与泛化能力的相关性

Fig. 6 Correlation between the model's attack rate and generalization ability

4.3 损失函数有效性比较

之前的实验证明了 DANet 是一个很好的代理模型，接下

来验证本文提出的损失函数修改是有效的。实验将使用DANet作为替代模型来证明语义消除对抗攻击方法具有更好的可迁移性。替代模型和受害模型都来自4.2节的实验。

对于基于梯度的攻击,实验将语义消除对抗攻击方法与FGSM和PGD进行比较。所有攻击方法都基于相同的约束:扰动的$\varepsilon=0.05$.对于基于GAN的攻击,将advGAN和Frequency-aware GAN(以下简称FA-GAN)训练150个周期并获取最终的对抗样本和语义消除对抗攻击方法比较。实验结果如表4所列,语义消除对抗攻击方法生成的对抗样本善于探索更多特征,因此可以更好地隐藏扰动。由于这些扰动更难检测到,因此它更真实地反映了模型对语义特征已被消除的图像的脆弱性。

表4 基于梯度的消除语义攻击的对照试验

Table 4 Controlled experiment on semantic attack based on gradient elimination

Dataset	Method	MVSS-Net	PSCC-Net	RRU-Net	SSIM
CASIAv2	Origin	0.8598	0.8363	0.8912	—
	FGSM	0.5574	0.4619	0.6583	0.8697
	PGD	0.6230	0.6227	0.6101	0.9146
	advGAN	0.6008	0.7323	0.7791	0.9504
	FA-GAN	0.4499	0.7086	0.6945	**0.9542**
	Ours	**0.2689**	**0.4245**	**0.5680**	0.9525
CASIAv1	Origin	0.4720	0.4215	0.3416	—
	FGSM	0.2344	0.3575	0.2536	0.8931
	PGD	0.2545	0.2899	0.2507	0.9271
	advGAN	0.2592	0.4011	0.2843	0.9442
	FA-GAN	0.2224	0.3648	0.2520	0.9501
	Ours	**0.2087**	**0.2326**	**0.2026**	**0.9588**

结束语 本文提出了一种用于攻击图像篡改定位模型的语义消除对抗攻击方法。具体地,提出了利用语义分割模型DANet作为图像篡改定位的替代模型,并提出了一种用于消除语义特征的损失函数。基于损失修正,对抗攻击可以更好地暴露模型的弱点。此外,由于本文方法是基于消除图像的语义特征,因此对抗样本也反映了图像篡改定位模型是否真正关注图像的篡改痕迹而不是只关注被篡改部分的语义。这种攻击方法迁移性强,可以攻击较多的现有模型。语义消除对抗攻击可以用于发现模型的弱点并对其进行设计改进,这是非常有意义的,值得进一步深入研究。

参 考 文 献

[1] BI X L, WEI Y, XIAO B, et al. RRU-Net: The ringed residual U-Net for image splicing forgery detection[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. 2019.

[2] LIU X H, LIU Y J, CHEN J, et al. PSCC-Net: Progressive spatio-channel correlation network for image manipulation detection and localization[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(11): 7505-7517.

[3] DONG C B, CHEN X R, HU R H, et al. MVSSNet: Multi-View Multi-Scale Supervised Networks for Image Manipulation Detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 45(3): 3539-3553.

[4] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples[J]. arXiv: 1412.6572, 2014.

[5] MADRY A, MAKELOV A, SCHMIDT L, et al. Towards deep learning models resistant to adversarial attacks[J]. arXiv: 1706.06083, 2017.

[6] ROZSA A, ZHONG Z, BOULT T E. Adversarial attack on deep learning-based splice localization[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. 2020: 648-649.

[7] GRAGNANIELLO D, MARRA F, POGGI G, et al. Analysis of adversarial attacks against CNN-based image forgery detectors [C]// 2018 26th European Signal Processing Conference (EUSIPCO). IEEE, 2018: 967-971.

[8] ZHU P, OSADA G, KATAOKA H, et al. Frequency-aware GAN for adversarial manipulation generation[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023: 4315-4324.

[9] MATTHEW D Z, FERGUS R. Visualizing and understanding convolutional networks[C]// Computer Vision — ECCV 2014: 13th European Conference, Zurich, Switzerland, Part I 13. Springer, 2014: 818-833.

[10] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]// Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017: 6000-6010.

[11] ZHANG Z Y, QIAN Y, ZHAO Y X, Noise and edge based dual branch image manipulation detection[C]// Proceedings of the 2023 4th International Conference on Computing, Networks and Internet of Things. 2023: 963-968.

[12] MA X C, DU B, JIANG Z H, et al. IML-ViT: Benchmarking Image Manipulation Localization by Vision Transformer[J]. arXiv: 2307.14863, 2023.

[13] TRIARIDIS K, MEZARIS V. Exploring Multi-Modal Fusion for Image Manipulation Detection and Localization[C]// Proceedings of the 30th International Conference on MultiMedia Modeling (MMM 2024). 2024.

[14] GUO X, LIU X H, REN Z Y, et al. Hierarchical Fine-Grained Image Forgery Detection and Localization[C]// CVPR. 2023.

[15] KURAKIN A, GOODFELLOW I, BENGIO S. Adversarial machine learning at scale[J]. arXiv: 1611.01236, 2016.

[16] MADRY A, MAKELOV A, SCHMIDT L, et al. Towards deep learning models resistant to adversarial attacks[J]. arXiv: 1706.06083, 2017.

[17] PAPERNOT N, MCDANIEL P, WU X, et al. Distillation as a defense to adversarial perturbations against deep neural networks[C]// 2016 IEEE Symposium on Security and Privacy (SP). IEEE, 2016: 582-597.

[18] PAPERNOT N, MCDANIEL P, JHA S, et al. The limitations of deep learning in adversarial settings[C]// 2016 IEEE European symposium on security and privacy (EuroS&P). IEEE, 2016: 372-387.

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

[20] PAPERNOT N, MCDANIEL P, GOODFELLOW I, et al. Practical black-box attacks against machine learning[C]// Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security. 2017: 506-519.

[21] PAPERNOT N, MCDANIEL P, GOODFELLOW I. Transfera-

bility in machine learning: from phenomena to black-box attacks using adversarial samples[J]. arXiv: 1605.07277, 2016.

[22] CHEN S Z, HE Z B, SUN C J, et al. Universal adversarial attack on attention and the resulting dataset damagenet[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 44(4): 2188-2197.

[23] HUANG H, CHEN Z Y, CHEN H R, et al. T-sea: Transfer-based self-ensemble attack on object detection[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 20514-20523.

[24] ZHU Y, CHEN C F, YAN G, et al. ARNet: Adaptive attention and residual refinement network for copy-move forgery detection [J]. IEEE Transactions on Industrial Informatics, 2020, 16(10): 6714-6723.

[25] HU X F, ZHANG Z H, JIANG Z Y, et al. Span: Spatial pyramid attention network for image manipulation localization[C]// Computer Vision — ECCV 2020: 16th European Conference, Glasgow, UK, Part XXI 16. Springer, 2020: 312-328.

[26] FU J, LIU J, TIAN H J, et al. Dual attention network for scene segmentation[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 3146-3154.

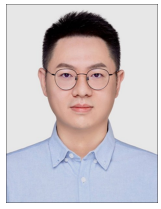
[27] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 770-778.

[28] GUILLARO F, COZZOLINO D, SUD A, et al. TruFor: Leveraging All-Round Clues for Trustworthy Image Forgery Detection and Localization[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2023: 20606-20615.

[29] WU H, CHEN Y, ZHOU J. Rethinking Image Forgery Detection via Contrastive Learning and Unsupervised Clustering[J]. arXiv: 2308.09307, 2023.

[30] KWON M J, NAM S H, YU I J, et al. Learning JPEG Compression Artifacts for Image Manipulation Detection and Localization [J]. International Journal of Computer Vision, 2022(8): 1875-1895.

JIANG Weihao, born in 2003, undergraduate. His main research interest is multimedia forensics and security.



LIU Bo, born in 1987, associate professor, supervisor, is a member of CCF (No. J4705M). His main research interest is multimedia forensics and security.