中图法分类号:TP309.7　　**文献标识码**:A　　**文章编号**:1006-8961(XXXX)XX-0001-18

论文引用格式:Ma Xuanbo, Zhang Shihao, Tian Huawei. Design of an integrated steganalysis model for adversarial steganography[J/OL]. Journal of Image and Graphics, XXXX:1-18. DOI: 10.11834/jig.260010. (马煊博, 张士豪, 田华伟. 针对对抗隐写的融合隐写分析模型设计[J/OL]. 中国图象图形学报,XXXX:1-18. DOI: 10.11834/jig.260010.)[DOI:10.11834/jig.260010]

针对对抗隐写的融合隐写分析模型设计

马煊博,张士豪*,田华伟

中国人民公安大学　信息网络安全学院,北京　100038

摘　要:**目的**　当前,基于深度学习的隐写分析方法相较于传统方法,虽在检测性能上展现出显著优势,却极易遭受对抗隐写方法的攻击。如何在隐写分析任务中协同发挥两类方法的优势,成为亟待解决的关键问题,基于此,提出一种融合式隐写分析框架。**方法**　以基于SRM(Spatial Rich Model)手工特征的传统隐写分析方法与深度学习隐写分析方法Ye-Net作为基学习器,通过集成学习对两者的判别输出进行融合;同时,构建基于对抗迁移网络的深度分类器,该分类器依托特征提取器与域判别器的对抗博弈过程,提取非对抗域与对抗域间可共享的域不变特征,实现了对抗域真值标签未知场景下的模型有效训练。此外,模型基于MLP(Multi-Layer Perceptron)构建偏离样本识别模块,有效抑制训练过程中出现的负迁移现象,稳定域分布对齐过程,进一步提升模型在对抗扰动环境下的跨域泛化能力。**结果**　实验结果表明,在不同嵌入率和不同强度的对抗隐写攻击下,所提融合隐写分析模型相较于SPAM(Subtractive Pixel Adjacency Matrix steganalysis)和SRM两种传统隐写分析模型错误率(Probability of Error,Pe)平均下降15.95%和6.06%,相较于深度学习隐写分析模型(Ye-Net、SRNet、LWENet)错误率平均下降了10.93%~19.50%,相较于针对对抗隐写方法的鲁棒性增强方法KDNFT(K-times Dropout Neighboring Feature Transformer)错误率平均下降5.90%,在对抗隐写场景下达成当前SOTA的隐写分析性能。**结论**　本文提出的融合式隐写分析框架,能够有效降低检测对抗样本隐写图像的综合错误率,为实现更加精准的高可靠隐写分析模型提供了新的可行路径。代码链接:https://doi.org/10.57760/sciencedb.j00240.00093

关键词:图像处理;隐写分析;深度学习;对抗隐写;集成学习;对抗迁移网络

Design of an integrated steganalysis model for adversarial steganography

Ma Xuanbo, Zhang Shihao*, Tian Huawei

College of Information and Cyber Security, People's Public Security University of China, Beijing 100038, China

Abstract: **Objective** At present, deep learning-based steganalysis methods generally outperform traditional handcrafted-feature approaches in conventional settings; however, under white-box adversarial attacks constructed by exploiting model gradients, their ability to detect adversarially generated stego images often degrades substantially. Meanwhile, the adversarial perturbations introduced to deceive deep neural networks may induce non-natural local pixel variations or abnormal neighborhood dependencies in stego images, thereby disrupting high-order statistical regularities. Such artifacts make adversarial stego images more detectable by traditional steganalysis methods that rely on handcrafted features and statistical analysis. Therefore, a key problem is how to effectively exploit the complementary strengths of deep learning-based and traditional approaches within a unified steganalysis framework. In addition, existing adversarial-training strategies typically

收稿日期:2026-01-06;**修回日期**:2026-03-26

*通信作者:张士豪,通信作者,男,硕士,讲师,主要研究方向为信息隐藏、社会网络分析等。E-mail:zhangshihao@ppsuc.edu.cn

基金项目:中央高校基本科研业务费(项目编号:2022JKF02020)

Supported by: the Fundamental Research Funds for the Central Universities

require a large number of adversarial stego images with ground-truth labels, whereas in real-world deployments such images are difficult to obtain and their labels are often unavailable. Motivated by these practical constraints, this paper proposes an integrated steganalysis framework designed for adversarial steganography scenarios. **Method** First, we employ a traditional steganalysis method based on SRM (Spatial Rich Model) handcrafted features and a deep learning-based steganalysis method, Ye-Net, as heterogeneous base learners. For an input sample, both detectors output the posterior probability that the sample is a stego image, and their outputs are integrated through an ensemble strategy to reduce the performance volatility of any single detector under adversarial perturbations. Since the two detectors exhibit markedly different output scales and calibration characteristics, we introduce a sigmoid-based alignment with a normalization intensity factor to match their output distributions and map them into a comparable probability space. After this calibration, we obtain a probability-level feature representation that serves as the input to the subsequent classifier. At the classifier level, we construct a deep classifier based on DANN (Domain-Adversarial Neural Network). The classifier consists of a feature extractor, a label predictor, and a domain classifier. The feature extractor and the domain classifier are connected via a gradient reversal layer and engage in an adversarial game: the feature extractor is trained not only to minimize the label predictor loss, but also to suppress domain separability, thereby learning domain-invariant features that are transferable across the non-adversarial domain and the adversarial stego domain. The learned features are then fed into the label predictor to produce the final label. This design enables effective training in scenarios where ground-truth labels in the adversarial stego domain are unknown or unavailable. In addition, we observe that in adversarial steganography settings a subset of adversarial stego images may undergo excessively large shifts in the feature space. During domain-adversarial alignment, such samples can be misleadingly pulled toward the cover region, which triggers negative transfer and deteriorates training effectiveness. To address this issue, we introduce a deviated-sample identification module, implemented with a multi-layer perceptron, before domain-adversarial training. The module identifies and filters out target-domain stego images with overly strong adversarial deviations, thereby mitigating negative transfer, stabilizing domain alignment, and further improving cross-domain generalization under adversarial perturbations. **Result** To evaluate robustness under different adversarial intensities, this paper constructs multiple test sets by mixing adversarial stego images and non-adversarial stego images at varying ratios. The experimental results reveal a distinct performance divergence among baseline detectors as the adversarial ratio increases. Deep learning-based models (including Ye-Net, SRNet, and LWENet) demonstrate superior detection accuracy in non-adversarial scenarios; however, their overall detection error Pe (Probability of Error) deteriorates significantly as the proportion of adversarial stego images rises, confirming that deep discriminative features are highly vulnerable to targeted gradient-based perturbations. Conversely, traditional methods based on handcrafted features (such as SRM and SPAM) exhibit an inverse trend, where error rates decrease or stabilize under high-intensity adversarial steganography conditions, indicating that statistical residual features are more sensitive to the abnormal artifacts introduced by adversarial perturbations. Leveraging the complementary nature of these characteristics, the proposed fusion-based steganalysis model consistently maintains the lowest and most stable detection error across most mixing ratios, outperforming even the specialized adversarial defense model, KDNFT. Experimental results demonstrate that under adversarial attacks of varying intensities, the proposed model reduces the average error rate by 15.95% and 6.06% compared to traditional models (SPAM and SRM, respectively), by 10.93% to 19.50% compared to deep learning models (Ye-Net, SRNet, and LWENet), and by 5.90% compared to the robustness-enhanced method KDNFT, achieving state-of-the-art (SOTA) performance in adversarial steganography scenarios. **Conclusion** This paper presents a fusion-based steganalysis framework that achieves stable and effective cross-domain alignment without requiring ground-truth labels in the adversarial stego domain. By integrating SRM and Ye-Net within an ensemble representation, aligning heterogeneous outputs via normalization intensity factors, and adopting a domain-adversarial transfer classifier enhanced with a deviated-sample filtering mechanism implemented with a multi-layer perceptron, the proposed approach substantially improves robustness and generalization under adversarial steganography conditions. The framework provides a practical and promising pathway toward more accurate and highly reliable steganalysis systems in adversarial environments.

Key words: image processing; steganalysis; deep learning; adversarial steganography; ensemble learning; domain-adversarial neural network

0 引 言

图像隐写作为隐写术的重要分支，其核心思想是利用特定编码方式将秘密信息嵌入载体图像中，使生成的载密图像在视觉上与原始载体几乎不可区分，从而为通信双方建立隐蔽信道。与之对抗的图像隐写分析旨在判定图像中是否存在秘密信息，以阻断潜在的隐蔽通信(龙玲慧等，2026)。现有通用型隐写分析方法大体可分为两类，一类是以SPAM(Pevný等，2010)、SRM(Fridrich等，2012)等为代表的传统方法，通常在残差域构造高维统计特征，并结合SVM(Support Vector Machine)(Suykens等，1999)或FLD(Fisher Linear Discriminant)(Li等，2014)等机器学习分类器进行判别，另一类是以Xu-Net(Xu等，2016)、Ye-Net(Ye等，2017)等为代表的深度学习方法，依托神经网络从数据中自动学习判别性特征，以减少人工特征工程对性能的限制。其中，Xu-Net(Xu等，2016)提出了端到端卷积神经网络用于隐写分析，从原始图像直接学习判别特征；Ye-Net(Ye等，2017)在网络前端引入面向残差信号的处理与结构约束，增强模型对微弱嵌入扰动的响应能力；SRNet(Boroumand等，2019)通过更深的网络层级与更强的特征表达能力学习复杂的隐写判别模式，在大规模数据条件下展现出更优的检测性能；LWENet(Weng等，2022)从轻量化建模角度出发，通过引入高效残差特征提取与参数压缩策略，在显著降低模型复杂度的同时保持了较强的隐写判别能力。总体而言，基于深度学习的隐写分析模型在传统隐写算法上已经取得了较高的检测性能，在数据量充足的条件下，深度学习方法相较传统特征分析框架通常能够获得更高的检测性能(陈君夫等，2021)。

近年来，深度神经网络被发现容易受到对抗隐写方法的攻击。由于深度学习模型在模型推理的过程中通常具有可微性，其反向传播梯度容易被基于梯度的白盒攻击所利用，从而定向生成对抗隐写图像以欺骗深度学习隐写分析模型。基于这一脆弱性，研究者提出了一类被称作对抗性隐写的新式方法，其核心在于利用模型梯度信息在像素层面施加幅度极小且方向敏感的扰动，使已训练完成的隐写分析模型以较高置信度做出错误判决。根据对抗扰动与嵌入流程的耦合方式，这类方法可分为两类：一类为嵌入无关型，即通过目标隐写分析器梯度信息在常规隐写嵌入前后单独施加对抗噪声，典型代表包括ADS(Adversarial Distortion-based Steganography)(Zhang等，2018)、SPS-ENH(Steganography with Perturbation and Selection - Enhanced)(Qin等，2021)和CAAS(Channel Attention Adversarial Steganography)(Sharma等，2023)；另一类为嵌入相关型，在嵌入环节引入梯度引导，让对抗优化直接参与载荷分配，如AEN(Adversarial Embedding Network)(Ma等，2019)、ADV-EMB(Adversarial Embedding)(Tang等，2019)、MAE(Mutual-information-based Adversarial Embedding)(Liu等，2021)以及多重对抗和通道注意力隐写(马宾等，2024)。总体而言，这类对抗隐写图像可被视作沿梯度方向构造的定向干扰，随着模型复杂度的提升与决策边界的细化，这种定向扰动更易贴合模型判别边界，从而增强其欺骗效果。

在对抗隐写场景中，攻击者仅对载密图像施加对抗扰动，攻击目标是诱导隐写分析器将载密图像误判为载体图像，从而实现隐写通信的隐蔽性增强。尽管基于深度学习的隐写分析模型在常规场景下通常优于传统方法，但当隐写数据受到对抗扰动时，其鲁棒性会显著降低。与此同时，对抗扰动在统计层面可能引入与隐写残差分布不一致的异常模式，使得不依赖梯度优化的传统空域隐写分析模型反而更容易识别对抗隐写图像。对于深度学习隐写分析模型，常见防御策略之一是对抗训练，即在训练集中显式引入特定算法产生的对抗隐写图像以促使模型学习相应特征。Goodfellow等人(2014)的研究解释了对抗样本的原理，并验证了对抗训练可以显著提升鲁棒性。Jawad等人(2025)采用对抗训练将特定扰动分布显式纳入训练过程，使模型在对抗分布下保持更稳定的判别性能。然而，由于非对抗隐写图像与对抗隐写图像存在显著的分布差异，在对抗博弈的环境下，系统的总体误差相较常规非对抗情形会有所上升。Lin等人的研究(2024)指出，对抗训练可能会降低深度学习模型在常规样本上的判别能力。

为进一步提升鲁棒性，研究者提出了多条辅助防御策略，典型代表包括采样-补丁式检测(Qin等，2022)、TStegNet(Hu等，2023)、RS-GAN和KDNFT(Lin等，2024)。总体而言，现有防御方法多依赖外部管线或复杂训练机制，对攻击方式、扰动强度或对

抗比例变化敏感。此外,在实际通信环境中,检测方通常难以事先获知对抗攻击的强度与比例,这种不确定性进一步限制了基于特定攻击分布的防御策略在复杂场景下的泛化能力。

为解决上述问题,本文引入域适应策略,对训练样本进行差异化处理。考虑到非对抗隐写图像与对抗隐写图像之间的特征差异,本文将非对抗隐写图像视为源域,将目标域构造为以一定比例混合非对抗隐写图像与对抗隐写图像的无真值样本集。同时,针对目标域中可能存在偏离过大,容易在域对抗训练中被错误对齐到载体图像区域的对抗隐写图像,本文进一步引入偏离样本识别机制,对该类图像进行过滤,以减少其对域适应训练的不良影响。在集成学习框架下,本文融合基于深度学习的隐写分析方法Ye-Net(Ye等,2017)和传统空域特征提取方法SRM(Fridrich等,2012)。在训练阶段,针对两类检测器的输出特性,分别使用Ye-Net与SRM训练多个基学习器,为缓解不同检测器输出尺度不一致的问题,进一步通过归一化强度因子将基学习器输出映射为统一维度的概率特征表示。在分类器设计方面,本文基于DANN(Domain-adversarial neural network)架构(Ganin等,2017),构建改进型隐写分析分类器,通过特征提取器与域分类器的对抗学习,在偏离样本过滤的基础上学习跨域可迁移特征,在不牺牲对非对抗隐写图像判别力的前提下,有效提升对对抗隐写图像的检测鲁棒性,从而增强模型的适用性与泛化能力。相比现有方法,本方法基于分布对齐与信息互补策略,对攻击细节和参数更不敏感;由于端到端的梯度传导在偏离样本识别机制处被截断,因此在面对基于梯度的攻击时具有较强鲁棒性,同时能够在不同对抗攻击的强度和比例下保持稳定的隐写分析性能。

本文的贡献主要有如下几点:

1)针对非对抗隐写图像与对抗隐写图像之间的特征差异,首次引入DANN域对抗学习框架,将非对抗隐写图像建模为源域,并将由非对抗隐写图像与对抗隐写图像按未知比例混合形成的无标注样本集建模为目标域,在不依赖目标域真值、也无需预先获知对抗比例的条件下,通过特征提取器与域分类器的对抗学习获得对任务判别保持敏感且对域偏移相对不敏感的可迁移特征,以缓解两域分布偏移对检测性能的影响。

2)提出面向目标域的偏离样本识别与过滤机制,以抑制与对抗训练中的负迁移。针对目标域中偏离过大,在域对抗训练过程中容易错误对齐到载体图像区域的对抗隐写图像,本文构建MLP(Multi-Layer Perceptron)过滤器,在训练阶段对这类图像进行过滤,从而减少其对域适应过程的干扰,提升训练稳定性与对抗场景下的鲁棒性

3)实现了面向对抗场景的异构检测器融合表征。在集成学习框架下,通过归一化强度因子对Ye-Net与SRM的输出进行统一校准与融合,使模型在保持Ye-Net对常规隐写样本较强判别能力的同时,引入SRM对对抗扰动更具稳健性的互补优势;从而兼顾常规检测性能与对抗环境下的鲁棒性,并在一定程度上提升系统对基于梯度白盒攻击的抗利用能力。

1 模型设计

1.1 威胁模型

根据密码学中的Kerckhoffs原则(Petitcolas等,1883),安全系统的可靠性应依赖于密钥的保密性,而非系统设计细节的隐蔽性。该原则在隐写与隐写分析的对抗框架中同样具有指导意义:隐写分析方已知目标隐写算法,并据此在训练阶段使用随机密钥合成带真值标注的干净样本用于监督学习。

然而,在真实对抗环境中,隐写分析方通常无法主动生成对抗隐写图像,也难以获知攻击策略及其强度;同时,实际截获的数据往往缺乏真值标签,对抗隐写图像在截获图像中的占比亦不可得。基于此,本文进一步假设:除可合成的干净有标签数据外,隐写分析方仅能获得一个来源于真实通信环境的无真值混合样本集,其中可能包含对抗隐写图像与非对抗隐写图像,其具体比例与攻击强度均未知。

1.2 总体概述

在隐写分析器面对未知样本时,若直接采用常规对抗训练方法,模型往往会对某一类对抗扰动或攻击模式产生过拟合,从而削弱其对非对抗隐写图像的判别能力。为避免对抗鲁棒性和对于非对抗隐写图像的检测性能之间的失衡,本文的目标是在保持非对抗隐写图像检测能力的前提下,进一步提升模型对对抗隐写图像的检测效果与泛化能力。

在特征构造方面,考虑到仅由非对抗隐写图像

训练得到的基学习器，在处理非对抗隐写图像与对抗隐写图像时会呈现出不同的特征响应与输出分布，本文将基学习器在两类数据下的预测概率输出作为融合特征，输入至后续分类器进行训练。具体而言，本文采用集成学习策略，以Ye-Net与SRM为两条基线构建多个集成学习器；所有基学习器仅使用非对抗隐写图像进行训练，以确保其预测结果能够呈现出非对抗隐写图像与对抗隐写图像的差异。

与此同时，为利用到真实对抗环境中可获得但无真值标注的数据，本文引入域对抗学习框架。定义基学习器训练集为$\boldsymbol{X}_B$，其中包含非对抗条件下的载体图像及其对应的载密图像。进一步地，将与$\boldsymbol{X}_B$不重叠的另一部分非对抗隐写图像视为源域训练集$\boldsymbol{X}_S$，将来自真实环境的无真值混合样本集作为目标域训练集$\boldsymbol{X}_T$，其样本数分别为N_S和N_T。其中，$\boldsymbol{X}_S$由载体图像和载密图像构成，且具备真值标注，而对于$\boldsymbol{X}_T$而言，其内部是否包含对抗隐写图像、对抗隐写图像占比以及攻击强度均未知。

本文所提方法的总体框架如图1所示。首先，基于$\boldsymbol{X}_B$训练多组集成学习器，包括Ye-Net分支与SRM分支。然后，利用训练好的集成学习器分别对$\boldsymbol{X}_S$和$\boldsymbol{X}_T$进行分类，提取其输出的预测概率，并将其拼接形成融合特征矩阵。由于Ye-Net与SRM的输出尺度存在差异，本文对融合特征进行归一化处理，并引入归一化强度因子以实现特征空间对齐。

在域对抗训练阶段，本文首先利用源域数据在Ye-Net与SRM的融合特征空间中训练MLP过滤器，用于评估样本与源域判别边界的一致性，并识别出偏离判别边界过大的异常样本，随后进行DANN训练时，仅使用Ye-Net特征作为输入，并在每轮训练中借助该过滤器从目标域$\boldsymbol{X}_T$中剔除差异过大的样本，以避免其被错误的对齐到载体图像区域，对DANN网络的训练产生负面影响。DANN分类器由特征提取器G_f、标签分类器G_y和域分类器G_d三部分组成，通过G_f与G_d的对抗博弈，网络能够学习到跨域可迁移特征。通过上述设计，DANN分类器能够在尽量维持对干净样本判别能力的同时，更稳健地学习目标域中的可迁移信息，从而提升对对抗隐写图像的识别与泛化性能。

在模型推理阶段，将待检测的未知样本集输入集成学习器得到预测概率特征；其中Ye-Net特征被送入DANN分支输出判别结果，SRM特征用于形成基于SRM的投票结果，最终将DANN输出与SRM投票结果进行加权融合，得到样本的最终分类标签，权重取决于样本是否被MLP过滤器识别为异常样本，对于异常样本，SRM权重会高于DANN权重。

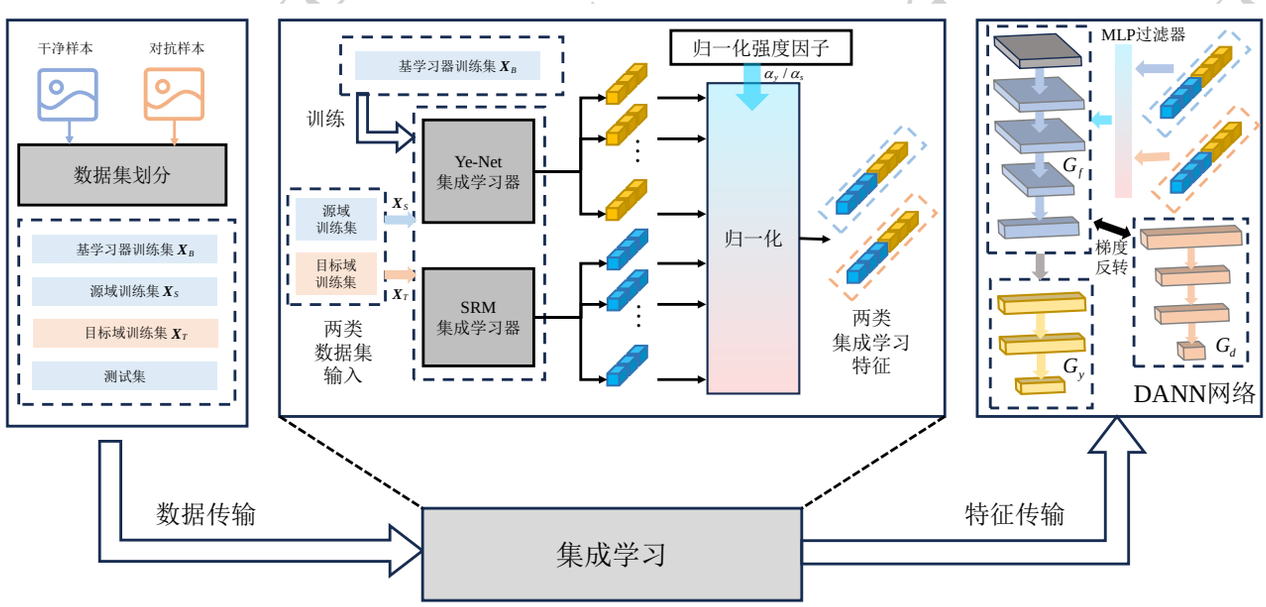


图1　总体框架图

Fig. 1　Overall framework diagram

1.3 集成学习

为获得高泛化性，本文采用集成学习分别构建Ye-Net与SRM两类基学习器集合，并将其输出统一转化为可融合的概率表示。对每一类检测器，本文使用Bootstrap重采样从基学习器训练集$\boldsymbol{X}_B$中构造K个训练子集，并在各子集上分别训练得到K个Ye-Net基学习器与K个SRM基学习器，以提升集成的多样性与泛化能力。记第k个Ye-Net基学习器为$g_y^{(k)}(\cdot)$，第k个SRM基学习器为$g_S^{(k)}(\cdot)$，式中$k=1,2,\cdots,K$。图2展示了模型数据集划分与集成学习部分的流程。

对于任意输入样本$\boldsymbol{x}$，Ye-Net基学习器的分类头采用二类SoftMax，因此可输出检测结果为载体图像或载密图像两类的后验概率向量：

$$g_y^{(k)}(\boldsymbol{x})=\left(\tilde{p}_{y,0}^{(k)}(\boldsymbol{x}),\tilde{p}_{y,1}^{(k)}(\boldsymbol{x})\right) \tag{3}$$

式中类别0表示载体图像，类别1表示载密图像。两类后验概率向量满足：

$$\tilde{p}_{y,0}^{(k)}(\boldsymbol{x})+\tilde{p}_{y,1}^{(k)}(\boldsymbol{x})=1 \tag{4}$$

直观上，$\tilde{p}_{y,1}^{(k)}(\mathbf{x})$越大，基学习器越倾向于将样本判为载密图像。然而，在与SRM进行融合时，直接使用原始SoftMax概率往往会出现两个问题：其一，不同基学习器的输出尺度与分布可能不一致，导致融合时某些基学习器的后验概率分布具有不同的熵水平与峰度特征，低熵输出更易在融合中占据主导，从而削弱其他学习器的互补信息；其二，在域偏移或对抗扰动存在的情况下，许多样本的SoftMax输出可能集中在0.5附近，难以体现样本间的相对差异。为此，本文引入归一化强度因子α_y，对Ye-Net的预测概率进行非线性校准：

$$p_y^{(k)}(\boldsymbol{x})=\sigma\left(\alpha_y\cdot\left(\tilde{p}_{y,1}^{(k)}(\boldsymbol{x})-\tilde{p}_{y,0}^{(k)}(\boldsymbol{x})\right)\right) \tag{5}$$

式中$\sigma(\cdot)$为Sigmoid函数：

$$\sigma(z)=\frac{1}{1+e^{-z}} \tag{6}$$

与Ye-Net不同，SRM分支首先提取高维SRM特征向量$\Phi(\boldsymbol{x})$，再使用传统分类器完成判别。本文在SRM基学习器中采用FLD集成框架，对第k个SRM基学习器，通过随机子空间策略构造M个特征子空间，并在各子空间上训练M个FLD分类器$\left\{f^{(k,m)}(\cdot)\right\}$。对于样本$\boldsymbol{x}$，SRM判别分数由所有子分类器输出进行累加得到：

$$s_s^{(k)}(\boldsymbol{x})=\sum_{m=1}^{M}f^{(k,m)}\left(\Phi(\boldsymbol{x})\right) \tag{7}$$

分数$s_s^{(k)}(\mathbf{x})$的符号表示分类结果，正值和负值分别表示分类结果为载体图像或载密图像，其绝对值反映判别的置信程度。为与Ye-Net分支在同一概率尺度上融合，本文同样使用归一化强度因子α_s将该分数映射为概率：

$$p_s^{(k)}(\boldsymbol{x})=\sigma\left(\alpha_s\cdot s_s^{(k)}(\boldsymbol{x})\right) \tag{8}$$

在获得两类基学习器的概率输出后，本文将其重组为统一的概率特征向量。对任意样本$\boldsymbol{x}$，定义预测概率向量：

$$\boldsymbol{p}(\boldsymbol{x})=\left[p_y^{(i)}(\boldsymbol{x}),\cdots,p_y^{(K)}(\boldsymbol{x}),p_s^{(i)}(\boldsymbol{x}),\cdots,p_s^{(K)}(\boldsymbol{x})\right]\in\mathbf{R}^{2K} \tag{9}$$

进一步地，将源域样本集$\boldsymbol{X}_S$和目标域样本集$\boldsymbol{X}_T$分别投入集成学习器，然后将得到的预测概率向量按行堆叠，得到对应的预测概率矩阵：

$$\boldsymbol{P}_S=\begin{bmatrix}\boldsymbol{p}(\boldsymbol{x}_1)\\ \vdots\\ \boldsymbol{p}(\boldsymbol{x}_{N_S})\end{bmatrix}\in\mathbf{R}^{N_S\times 2K},\boldsymbol{P}_T=\begin{bmatrix}\boldsymbol{p}(\boldsymbol{x}_1)\\ \vdots\\ \boldsymbol{p}(\boldsymbol{x}_{N_T})\end{bmatrix}\in\mathbf{R}^{N_T\times 2K} \tag{10}$$

式中$\boldsymbol{P}_S$和$\boldsymbol{P}_T$分别作为后续DANN分类器阶段的输入特征表示。通过上述构造，本文将深度学习模型Ye-Net的输出与传统特征模型SRM的输出统一在同一概率空间内，从而为后续跨域学习与鲁棒融合提供稳定、可比且易于训练的输入。

1.4 DANN分类器结构

为实现对抗环境下更稳定的隐写判别，本文将上一节构造的源域预测概率矩阵$\boldsymbol{P}_S$和目标域预测概率矩阵$\boldsymbol{P}_T$作为输入，引入基于DANN架构的隐写分析分类器以提升跨域泛化能力，其主要由特征提取器$G_f(\cdot;\theta_f)$，标签分类器$G_y(\cdot;\theta_y)$和域分类器$G_d(\cdot;\theta_d)$组成，其中θ_f、θ_y和θ_d分别是特征提取器、标签分类器和域分类器的内部参数。需要强调的是，目标域样本用于模拟真实检测环境中截取到的无真值图像集合，其内部可能混入一定比例的对抗隐写图像，但训练阶段不需要目标域的类别真值，也无需预先获知对抗隐写图像比例。

在域对抗训练中，若目标域样本的分布与源域差异过大，部分对抗隐写图像在特征空间中可能更接近源域的载密图像区域，从而在对齐过程中被错误的对齐到载体图像簇。导致特征提取器学习到不合理的对齐方向，引发负迁移。为直观展示这类现象，本文对Ye-Net和SRM两类基学习器输出的载体

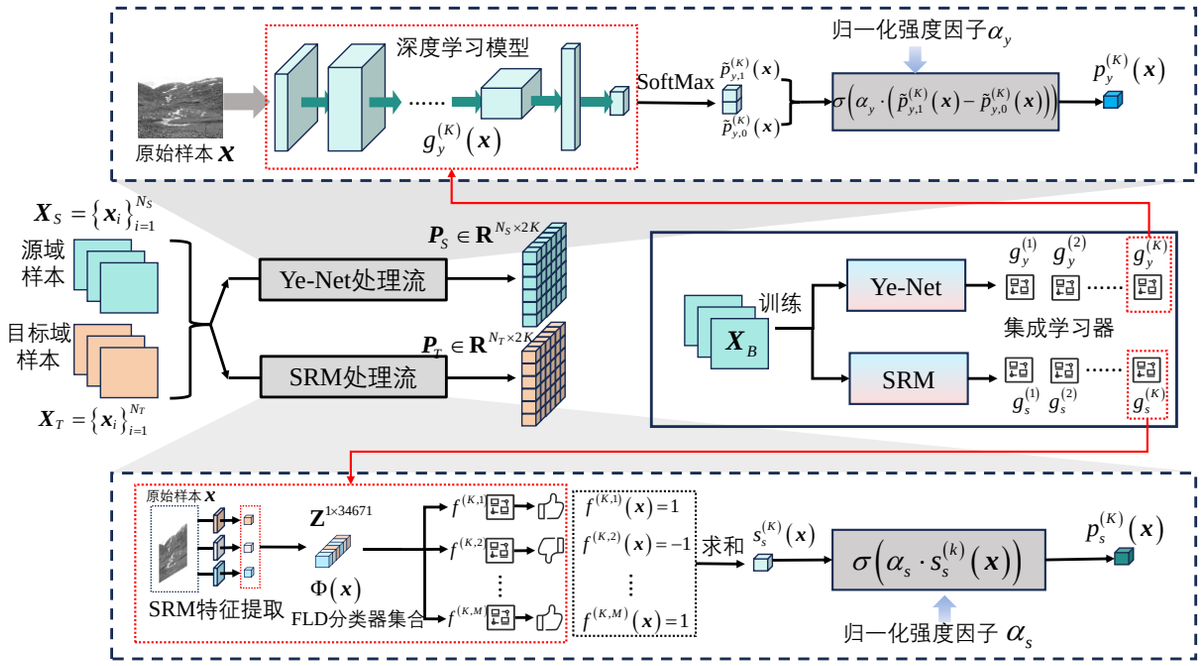


图2　集成学习图示

Fig. 2　Illustration of ensemble learning

图像(cover)、非对抗载密图像(stego)和对抗隐写图像(adv)的预测概率特征进行PCA降维可视化，结果如图3所示。可以观察到，相较SRM概率特征，Ye-Net概率特征在对抗隐写图像上更容易出现明显偏移；若直接使用Ye-Net特征进行无差别对齐，将更容易使偏移较大的对抗隐写图像干扰对齐过程，进而降低最终分类性能。

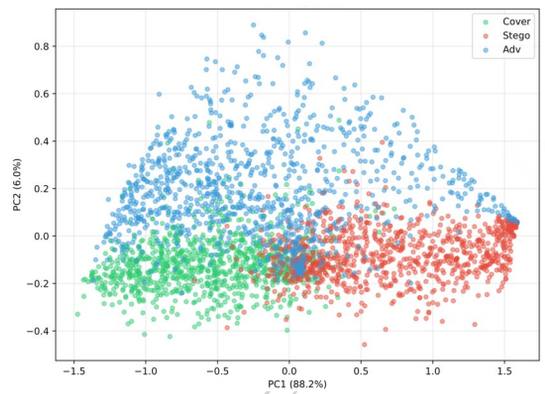
基于上述观察，本文在DANN分类器之前引入一个基于MLP的偏离样本识别模块，用于从目标域中筛除偏移过大的样本，以降低其对域对抗学习的干扰并缓解负迁移风险。

如图4所示，MLP过滤器由两个结构对称的判别头与一个阈值分类器组成，两个判别头分别处理Ye-Net和SRM两路概率特征向量，每个判别头均采用多层感知机结构，由五层全连接网络堆叠而成，各层采用ReLU激活函数，引入Dropout以抑制过拟合，并使用批归一化(Batch Normalization)以稳定训练过程。设传入图像为$\boldsymbol{x}$，判别头最终输出Ye-Net和SRM对样本$\boldsymbol{x}$的总体置信度，分别记为$p_y(\boldsymbol{x})$和$p_s(\boldsymbol{x})$，其中下标y,s分别对应Ye-Net分支和SRM分支，且两者均表示样本为载密图像的置信度。为刻画两路检测器在同一样本上的方向性分歧，本文构建方向性差值指标：

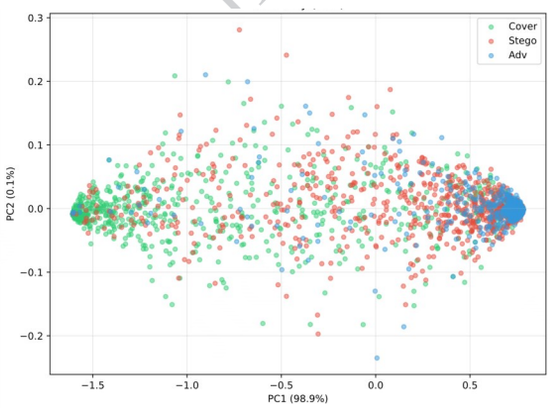
$$\Delta(\boldsymbol{x})=p_y(\boldsymbol{x})-p_s(\boldsymbol{x}) \tag{11}$$

当$\Delta(\boldsymbol{x})$显著为负时，意味着Ye-Net对分支样本的置信度明显低于SRM分支，即该样本在Ye-Net概率特征空间中更可能发生错误对齐到载体图像的偏离现象。此时若将其纳入域对抗训练，容易诱导特征提取器学习到不合理的对齐方向，从而对DANN训练产生反作用。本文设定过滤判定阈值τ作为门控判定准则，当$\Delta(\boldsymbol{x})<\tau$时，将样本判定为偏离样本。对于偏离样本，推理阶段将提高SRM分支的权重，训练阶段则将其直接丢弃，以减少其对域对抗训练的不利影响。MLP过滤器采用预训练方式得到参数。具体地，两个判别头分别以源域样本的Ye-Net概率特征与SRM概率特征为输入，并利用源域标签进行监督训练。

在此基础上，本文构建DANN分类器用于跨域判别。需要说明的是，DANN训练阶段仅使用源域与目标域的Ye-Net概率特征进行域对抗学习，而目标域特征在输入DANN前需先经过MLP过滤器剔除偏离样本。DANN分类器由三部分组成，特征提取器$G_f(\cdot;\theta_f)$能够将低维概率特征映射到更具判别性的表示空间，为后续标签分类与提供跨域特征，本文采用五层全连接网络对输入特征进行非线性变换，并在中间层引入批归一化，使用Dropout抑制过



（a） Ye-Net特征降维



（b） SRM特征降维

((a)Ye-Net dimensionality reduction;(b)SRM dimensionality reduction)

图3 PCA降维可视化结果

Fig. 3 PCA-based dimensionality reduction visualization results

拟合，激活函数采用ReLU。标签分类器$G_y(\cdot;\theta_y)$能够基于特征提取器输出的高维特征f，完成对载体图像和载密图像的判别，保证模型在两域上的检测性能。域判别器$G_d(\cdot;\theta_d)$用于预测样本来自源域还是目标域，并与特征提取器形成对抗关系。$G_d(\cdot;\theta_d)$由轻量全连接网络组成，输入为高维特征f，输出为域标签$\hat{d}$，设输入样本为$\boldsymbol{x}$，则

$$\hat{d}(\boldsymbol{x})=G_d\left(G_f\left(\boldsymbol{p}_y(\boldsymbol{x});\theta_f\right);\theta_d\right) \tag{12}$$

式中$\boldsymbol{p}_y(\boldsymbol{x})$为样本$\boldsymbol{x}$的Ye-Net概率特征。设标签分类器损失为$L_y$，域判别器损失为$L_d$，为实现对抗训练，在$G_f$与$G_d$之间插入梯度反转层(Gradient Reversal Layer，GRL)，记梯度反转算子为$R(\cdot)$，其输入为x，则其前向传播为恒等映射，而反向传播对梯度取负值：

$$R(x)=x \tag{13}$$

$$\frac{dR}{dx}=-\beta\boldsymbol{I} \tag{14}$$

其中$\boldsymbol{I}$为单位矩阵，β为梯度反转系数，控制了域对抗的强度。在DANN网络的训练过程中，定义三类损失函数：

$$L_y=\frac{1}{n}\sum_{i=1}^{n}L_y^i\left(G_y\left(G_f\left(\boldsymbol{p}_y(\boldsymbol{x}_i);\theta_f\right);\theta_y\right),y_i\right) \tag{15}$$

$$L_d^S=\frac{1}{n}\sum_{i=1}^{n}L_d^i\left(G_d\left(R\left(G_f\left(\boldsymbol{p}_y(\boldsymbol{x}_i);\theta_f\right)\right);\theta_d\right),d_i\right) \tag{16}$$

$$L_d^T=\frac{1}{n'}\sum_{i=n+1}^{N}L_d^i\left(G_d\left(R\left(G_f\left(\boldsymbol{p}_y(\boldsymbol{x}_i);\theta_f\right)\right);\theta_d\right),d_i\right) \tag{17}$$

式中L_y为标签分类损失，L_d^S和L_d^T分别为源域样本和目标域样本的域分类损失，n和n'分别为源域和目标域样本的数量，DANN分类器的成本函数为：

$$E=L_y-\lambda\left(L_d^S+L_d^T\right) \tag{18}$$

式中$\lambda>0$为域对抗权重系数。

在推理阶段，模型不再进行参数更新，而是对未知样本执行固定的推理流程，以输出最终分类结果，首先，将待检测样本$\boldsymbol{x}$输入已训练完成的集成学习器，得到预测概率矩阵$\boldsymbol{p}(\boldsymbol{x})$，其中包含两类概率特征向量$\boldsymbol{p}_y(\boldsymbol{x})$和$\boldsymbol{p}_s(\boldsymbol{x})$，并执行与训练阶段一致的归一化与对齐处理。随后，将$\boldsymbol{p}_y(\boldsymbol{x})$与$\boldsymbol{p}_s(\boldsymbol{x})$投入至MLP过滤器并计算$\Delta(\boldsymbol{x})$，判断样本是否为偏离样本，用于后续融合权重的自适应选择。对于DANN分支，将$\boldsymbol{p}_y(\boldsymbol{x})$输入由$G_f$和$G_y$组成的前馈网络，得到DANN分类置信度$p_D(\boldsymbol{x})$。对于SRM分支，本文对SRM概率特征向量取均值以获得稳定的分支置信度，并采用分段加权策略融合两分支输出：

$$p(\boldsymbol{x})=w_D p_D(\boldsymbol{x})+w_S\cdot\mathrm{mean}\left(\boldsymbol{p}_s(\boldsymbol{x})\right) \tag{19}$$

其权重(w_D,w_S)由偏离样本判定结果决定，当$\boldsymbol{x}$为偏离样本时，w_S需要大于w_D，以增强SRM的作用，当$\boldsymbol{x}$为非偏离样本时采用更均衡的权重分配。通过上述机制，对于偏离程度较强的对抗隐写图像，模型更多依赖对对抗扰动更具稳健性的SRM判别结果；对于偏离程度较弱的对抗隐写图像，则更多利用DANN学到的跨域可迁移特征进行判别，从而在未知攻击强度与比例条件下提升整体检测性能与稳定性。

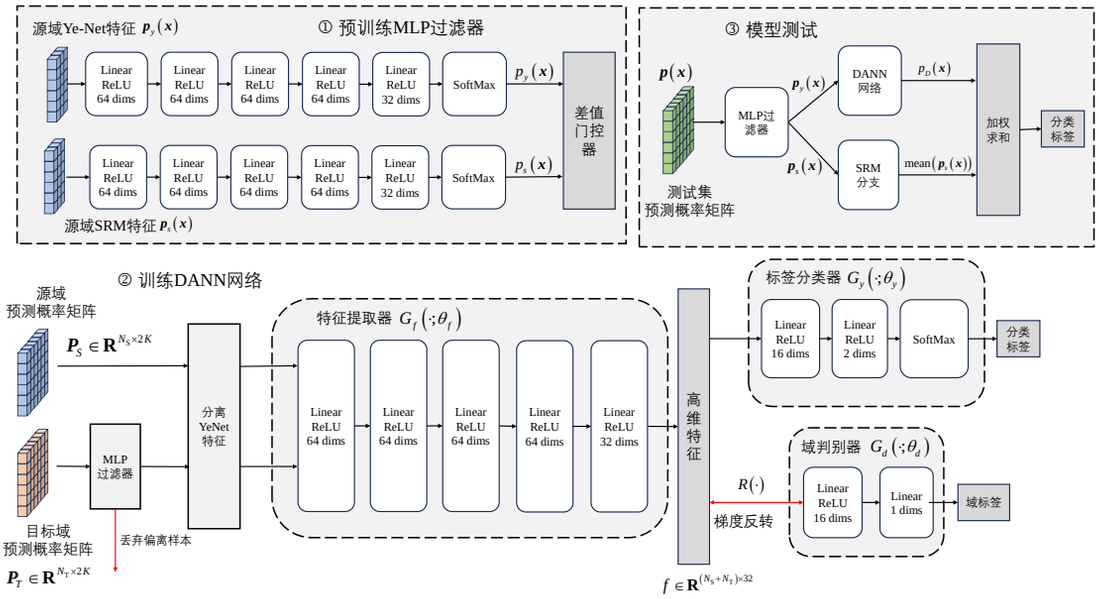


图4　DANN分类器结构图

Fig. 4　Architecture of the DANN classifier

2　实　验

2.1　实验设置

实验部分将展示模型在构建与训练过程中的参数敏感性及其纵向、横向对比效果，以及模型各模块的作用。实验基于BOSS base v1.01数据集(Bas等，2011)与BOWS2数据集(IEEE Signal Processing Society. 2025)进行，二者均为隐写与隐写分析领域所通用的数据集。BOSS base v1.01数据集包含10000幅512×512分辨率的灰度图像，BOWS2数据集同样由10000张512×512分辨率的灰度图像组成，作为BOSS base v1.01数据集的扩展。本文取前15000张图像作为基学习器训练集，后1000张图像作为测试集，剩余图像作为分类器训练集。

为模拟在实际对抗环境中隐写分析方获取到的混合样本集，对于分类器训练集，基于ADV-EMB(Tang等，2019)使用所有载体图像生成对抗隐写图像，并按照0%、25%、50%、75%、100%五种对抗比例与非对抗隐写图像相混合，生成多组混合样本集。测试集同样使用该混合样本集进行，以测试模型在对抗比例逐渐升高时的表现趋势。

集成学习阶段，所有基学习器的训练用子集均通过Bootstrap重采样从基学习器训练集$\mathbf{X}_B$中得到。由于Ye-Net基学习器需要划分训练集与验证集，本文在训练Ye-Net基学习器时，在其训练用子集内随机抽取80%的图像作为训练集，20%的图像作为验证集，使用Ye-Net和SRM两种基学习算法各训练了10个基学习器。同时，需要对Ye-Net与SRM两条分支产生的预测概率矩阵进行尺度对齐，本文对各分支输出分别采用带有归一化强度因子的Sigmoid函数进行归一化，并对所有保存的概率矩阵执行批量归一化后作为后续模型输入。

对于MLP过滤器，其分别以Ye-Net概率特征和SRM概率特征作为输入，每一层的Dropout比例为0.3，训练轮数为30，采用SGD优化器，学习率为0.01，动量优化参数(momentum)为0.9，权重衰减为0.0001，并使用StepLR学习率调度以提升收敛稳定性(step size=10，γ=0.5)。对于DANN网络的训练，设置训练轮数为100，批次大小为64，优化器同样采用SGD优化器，学习率为0.01，动量优化参数(momentum)为0.9，权重衰减为0.0001，同样采用StegLR调整学习率(step size=30，γ=0.5)。

实验指标方面，本文使用P_E值(Probability of Error，P_E)作为检测指标，判断是否为载密图像的阈值设置为0.5，其计算公式为：

$$P^{E0.5}=\frac{1}{2}\left(\frac{FP}{TN+FP}+\frac{FN}{TP+FN}\right) \tag{20}$$

式中，TP为正样本被预测为正样本的数量，FN为正

样本被预测为负样本的数量，TN为负样本被预测为负样本的数量，FP为负样本被预测为正样本的数量。$P^{E0.5}$值越低，表明隐写分析模型效果越好。

2.2 关键参数敏感性分析

在这一部分中，将对关键参数进行敏感性分析，主要包括Ye-Net分支与SRM分支的归一化强度因子、MLP过滤器阈值τ和偏离样本与非偏离样本的权重设置三方面内容。由于梯度反转系数β与域对抗权重系数λ共同控制域对抗强度，本文不再对二者的参数进行分析，而是参考Ganin等人的研究，设置$\lambda=1$，β使用经典的渐进式调度，使得其取值随着训练进度从0渐进至1。设p为当前训练轮数与总训练轮数的商，则：

$$\beta=\frac{2}{\left(1+e^{-10p}\right)-1} \tag{21}$$

关于Ye-Net分支和SRM分支的归一化强度因子α的取值，本文引入信息熵H作为归一化效果的度量指标，用以衡量归一化后概率分布的离散程度与有效区分能力。具体地，本文在离散候选集合上对Ye-Net分支和SRM分支的两类α的最优取值进行网格搜索，令$\alpha\in[0.001,0.4]$，步长为0.001，对每个候选α，将对应分支的输出经Sigmoid校准后计算信息熵，并以熵值大小作为选取依据，不同α下的信息熵变化曲线如图5所示：

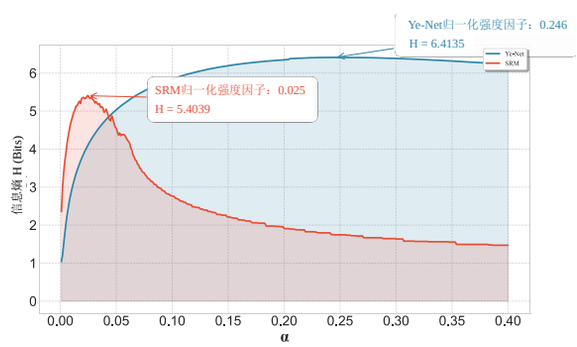


图5 基于信息熵的归一化强度因子选择

Fig. 5 Selection of the Normalization Intensity Factor Based on Information Entropy

根据结果，最终归一化强度因子分别设置为Ye-Net分支0.246，SRM分支0.025，能够使两分支的归一化结果具有更合适的分布形态，从而为后续融合与域适应训练提供稳定、可比的概率特征表示。

接着，关于MLP过滤器的阈值设置，首先在[-0.8,-0.1]区间以步长为0.1进行粗粒度的网格搜索，在每一个对抗比例下，记录模型的$p^{E0.5}$值，并计算不同对抗比例下的$p^{E0.5}$均值。根据搜索结果，最优阈值区间位置为[-0.6,-0.4]，随后，以0.001进行细粒度搜索，最终结果如图6所示：

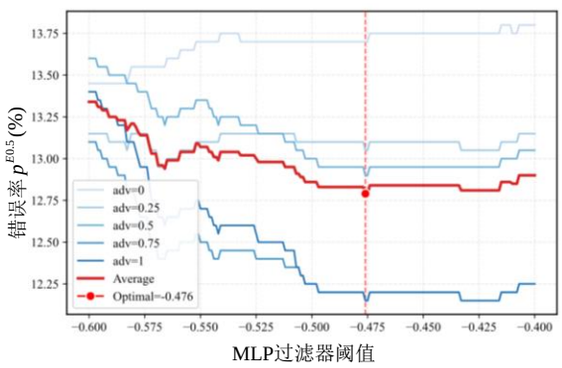


图6 MLP过滤器阈值分析

Fig. 6 Analysis of the MLP Filter Threshold

最后，关于偏离样本与非偏离样本的权重设置，采用二维网格搜索方法，在[0,1]区间内以步长0.1进行粗粒度定位。根据搜索结果，定位区间分别为[0.2,0.4]和[0.4,0.6]。然后，以步长0.001进行细粒度搜索，最终结果如图7所示。

从热力图可见，偏离样本权重存在一个最低下限，若权重低于此下限，会导致模型对对抗隐写图像的检测效果显著下降。进一步对不同对抗比例下偏离样本权重和非偏离样本权重对模型效果的影响进行分析，结果如图8、图9所示。

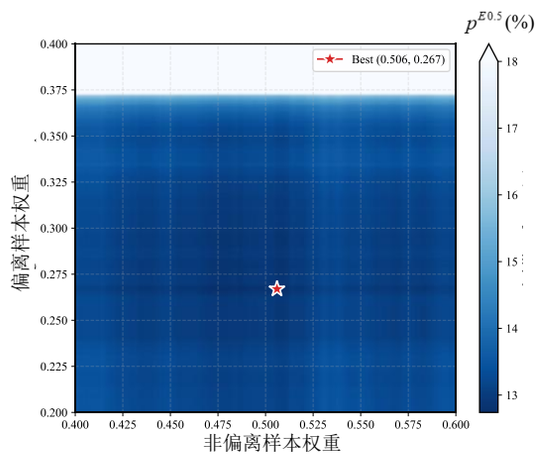


图7 偏离样本与非偏离样本权重取值联合热力图

Fig. 7 Joint Heatmap of Weights for Unknown and Known Samples

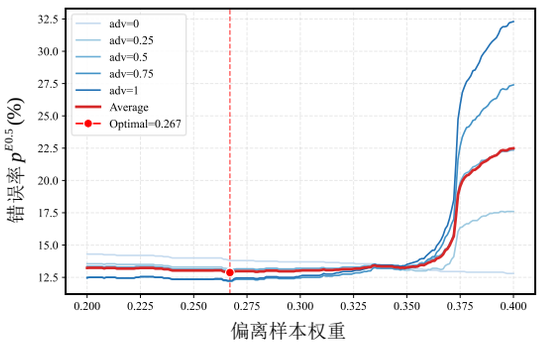


图8　偏离样本权重分析

Fig. 8　Analysis of the Weight for Unknown Samples

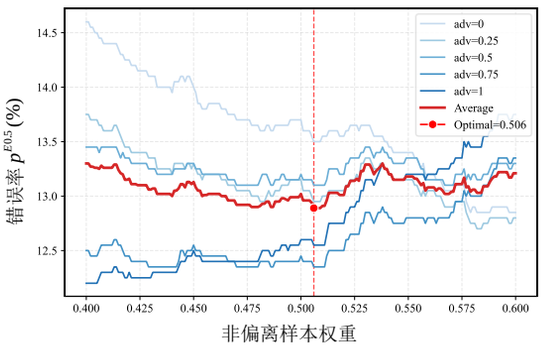


图9　非偏离样本权重分析

Fig. 9　Analysis of the Weight for Known Samples

进一步分析表明，偏离样本的权重取值对检测效果的影响较为显著，尤其当偏离样本权重接近0.375时，$p^{E0.5}$值出现明显的断崖式上升，且随着对抗强度提升（即对抗隐写图像所占比例的增加），上升幅度更加明显。这一现象表明，偏移样本过滤机制在对抗隐写图像的识别与检测中起到重要作用。

相较之下，非偏离样本的权重取值对于模型性能的影响主要体现在对抗比例的差异。高对抗比例下，$p^{E0.5}$随权重的增大而显著上升，在低对抗比例下权重增加则会导致$p^{E0.5}$值下降。这种差异导致取平均后，模型性能稳定在某一区间范围内，图7中呈现出较为明显的横向条纹，而非纵向条纹。

综上，偏离样本和非偏离样本的最优权重分别取五种对抗比例下的最优平均值，分别为0.267和0.506。该配置使得模型在面对不同对抗比例时，能够保持较为稳定且高效的检测性能。

2.3　融合模型测试结果

为了确保对比的广泛性与严谨性，实验选取了三类具有代表性的隐写分析方法作为基准：包括经典的传统手工特征方法SPAM（Pevný等，2010）、SRM（Fridrich等，2012），当前主流的深度学习隐写分析网络Ye-Net（Ye等，2017）、SRNet（Boroumand等，2019）和LWENet（Weng等，2022），以及针对对抗隐写的鲁棒性增强方法KDNFT（Lin等，2024）。目标隐写算法选用S-UNIWARD（Holub等，2013）和HILL（Li等，2014），二者均为自适应隐写算法，被广泛用于基准评测。对抗嵌入方法使用ADV-EMB（Tang等，2019），用于对载体图像进行对抗嵌入，该方法需要依托既定隐写算法生成对抗隐写图像。对抗载荷为0.2bpp、0.4bpp和0.6bpp。通过测试集载密图像中对抗隐写图像所占比例0%至100%的变化过程，对上述模型进行横向对比，旨在系统地分析各模型在面对不同程度对抗干扰时的性能演变趋势，实验结果见表1。

综合表1中实验结果可见，在嵌入率为0.2bpp与0.4bpp下，所提出的融合模型在大部分对抗比例下的$p^{E0.5}$平均值均位于首位，表现出显著优势。在0.6bpp条件下，其整体性能排名第二，$p^{E0.5}$平均值略高于KDNFT，仍保持较强的竞争力。针对以Ye-Net为攻击目标构造的对抗隐写图像，对深度学习隐写分析模型SRNet、LWENet也会起到对抗的效果，在不同的嵌入率下，$p^{E0.5}$值均会有不同程度的下降，说明了单一深度模型在跨模型对抗干扰下仍存在一定的脆弱性。对于传统方法SPAM，由于其主要基于相邻像素差分序列构建一阶马尔可夫链模型，并通过统计转移概率矩阵作为判别特征，梯度驱动的对抗攻击在扰动图像时往往会进一步破坏像素间的自然相关性。这种结构性扰动反而强化了差分统计特征的异常性，使得隐写信号更易被捕获。因此，随着对抗比例提升，SPAM的$p^{E0.5}$值反而呈现下降趋势。

相较之下，SRM对对抗隐写图像的检测效果会随着嵌入率的升高而有所下降。在无对抗（对抗比例为0%）和满对抗（对抗比例为100%）条件下，$p^{E0.5}$的差值在嵌入率为0.2bpp、0.4bpp和0.6bpp下分别为0.1031、0.0585和-0.0888，表明在高嵌入率叠加高对抗强度的极端环境中，传统高维特征模型同样会受到影响。由于融合模型的效果受制于两类基学习算法的表现，在高嵌入率高对抗环境下Ye-Net与SRM的$p^{E0.5}$值均出现一定程度的下降，导致在0.6bpp条件下的高对抗场景中，融合模型$p^{E0.5}$值略

表 1　S-UNIWARD隐写算法下不同对抗比例条件的$P^{E0.5}$值对比结果

Table 1　Comparison of $P^{E0.5}$ Values of Different Models under Various Adversarial Ratios for the S-UNIWARD Steganographic Algorithm

嵌入率/对抗比例	Ye-Net	SRNet	LWENet	SPAM	SRM	KDNFT	Ours
0.2bpp							
0%	**0.269 6**	0.472 8	0.299 5	0.445 4	0.366 1	0.359 7	0.276 0
25%	0.356 3	0.483 4	0.350 2	0.433 7	0.340 0	0.340 0	**0.260 5**
50%	0.439 2	0.495 7	0.399 4	0.421 1	0.316 5	0.320 7	**0.239 0**
75%	0.528 0	0.504 9	0.447 3	0.407 3	0.290 9	0.301 1	**0.219 5**
100%	0.611 9	0.515 8	0.497 3	0.395 7	0.263 0	0.282 4	**0.203 0**
平均	0.441 0	0.494 5	0.398 7	0.420 6	0.315 3	0.320 8	**0.239 6**
0.4bpp							
0%	0.141 9	0.266 8	**0.112 5**	0.351 0	0.200 5	0.139 4	0.135 5
25%	0.183 9	0.308 2	0.177 8	0.321 6	0.187 0	0.147 6	**0.127 0**
50%	0.223 2	0.345 5	0.243 6	0.293 2	0.174 0	0.161 1	**0.125 5**
75%	0.263 7	0.378 5	0.310 2	0.261 0	0.155 0	0.173 0	**0.118 0**
100%	0.301 7	0.390 2	0.376 0	0.234 2	0.142 0	0.182 2	**0.117 0**
平均	0.222 9	0.337 8	0.244 0	0.292 2	0.171 7	0.160 7	**0.124 2**
0.6bpp							
0%	0.104 0	0.159 8	0.110 3	0.260 2	0.133 6	0.131 9	**0.089 5**
25%	0.218 3	0.225 4	0.143 6	0.248 7	0.155 8	0.142 6	**0.129 0**
50%	0.331 5	0.292 8	0.178 5	0.234 9	0.177 0	**0.153 7**	0.176 0
75%	0.444 8	0.360 9	0.213 6	0.222 9	0.197 1	**0.167 9**	0.210 0
100%	0.558 6	0.429 7	0.244 9	0.212 5	0.222 4	**0.175 4**	0.253 0
平均	0.331 4	0.293 7	0.178 2	0.235 8	0.177 2	**0.154 3**	0.171 5

低于KDNFT。然而，从整体趋势来看，在绝大多数对抗比例与嵌入率组合下，融合模型均保持最优或次优表现，体现出良好的稳定性与泛化能力。

为进一步全面评估所提模型的综合性能与泛化能力，本文选用HILL隐写算法，在嵌入率为0.4 bpp的条件下对各对比算法进行测试，实验结果如表2所示：

表 2　HILL隐写算法下不同对抗比例条件的$P^{E0.5}$值对比结果

Table 2　Comparison of $P^{E0.5}$ Values of Different Models under Various Adversarial Ratios for the HILL Steganographic Algorithm

对抗比例	Ye-Net	SRNet	LWENet	SPAM	SRM	KDNFT	Ours
0%	**0.231 6**	0.381 5	0.273 8	0.467 7	0.401 1	0.382 6	0.262 0
25%	0.301 5	0.400 7	0.326 5	0.460 3	0.370 1	0.374 0	**0.244 0**
50%	0.374 1	0.419 9	0.382 3	0.454 9	0.346 3	0.365 7	**0.228 0**
75%	0.446 2	0.437 3	0.435 9	0.448 1	0.315 2	0.354 5	**0.214 0**
100%	0.518 5	0.456 3	0.487 5	0.441 5	0.283 1	0.347 5	**0.201 0**

在嵌入率为0.4bpp的条件下，对比S-UNIWARD与HILL两种隐写算法下的实验结果可以发现，在无对抗条件下，S-UNIWARD场景下各模型的$p^{E0.5}$值整体低于HILL场景，说明在相同嵌入率下，HILL算法对检测器而言相对更具挑战性。在对抗比例逐渐提升的过程中，两种算法呈现出不同的性能变化趋势。在S-UNIWARD场景下，随着对抗比例提升，深度模型的性能退化较为明显，而传统统计模型在该对抗比例下反而呈现一定的稳定甚至改善趋势。而在HILL场景下，深度模型的退化幅度更为平滑，传统模型的优势则不如在S-UNIWARD条件下明显，这表明不同的隐写算法会影响对抗扰动对统计结构的破坏方式。

进一步对比可以发现，所提出的融合隐写分析模型在S-UNIWARD与HILL两种隐写算法下均保持较优或最优表现，其$p^{E0.5}$平均值分别为0.1242和0.2010。且融合模型在对抗比例变化过程中波动幅度较小，尤其在中高对抗比例条件下，融合模型均能够有效抑制性能退化，表现出较强的算法无关性与泛化能力。进一步地，本文在S-UNIWARD算法0.4bpp场景中，通过调整判断载密图像的阈值，得到各模型的ROC曲线如图10所示。

从图10(f)可以看出，融合模型取得了全场最高的平均AUC值(0.949)，超越了针对对抗隐写的鲁棒性增强模型KDNFT(0.941)，并显著优于SRM(0.916)和Ye-Net(0.817)两大基准模型，这表明融合模型具备了跨越不同安全环境的广义鲁棒性，在不同的对抗比例下，模型均能维持较高的综合判别能力。

值得注意的是，在图10(d)和图10(e)的高对抗场景中，融合模型的ROC曲线出现了TPR快速下降但FPR几乎保持不变的情况，这是由于融合模型最终的输出结果并不只依赖DANN网络，还依赖SRM的投票结果，导致模型的最终预测分数集中在若干特定值附近，而非连续分布。进一步地，以对抗比例为50%为例，正样本数量最多的预测分数区间为[0.6,0.65]，而不是连续分布中预测分数越接近1正样本数量越多的情况，对比预测分数区间[0.6,0.65]和[0.65,0.7]，正样本数量快速下降，但负样本数量却变化不明显，这就导致TPR直线下降而FPR几乎没有变化的情况。在实际的判断中，判断是否为载密图像的阈值常取0.5，因此这一情况不会过于影响融合模型的判别效果。

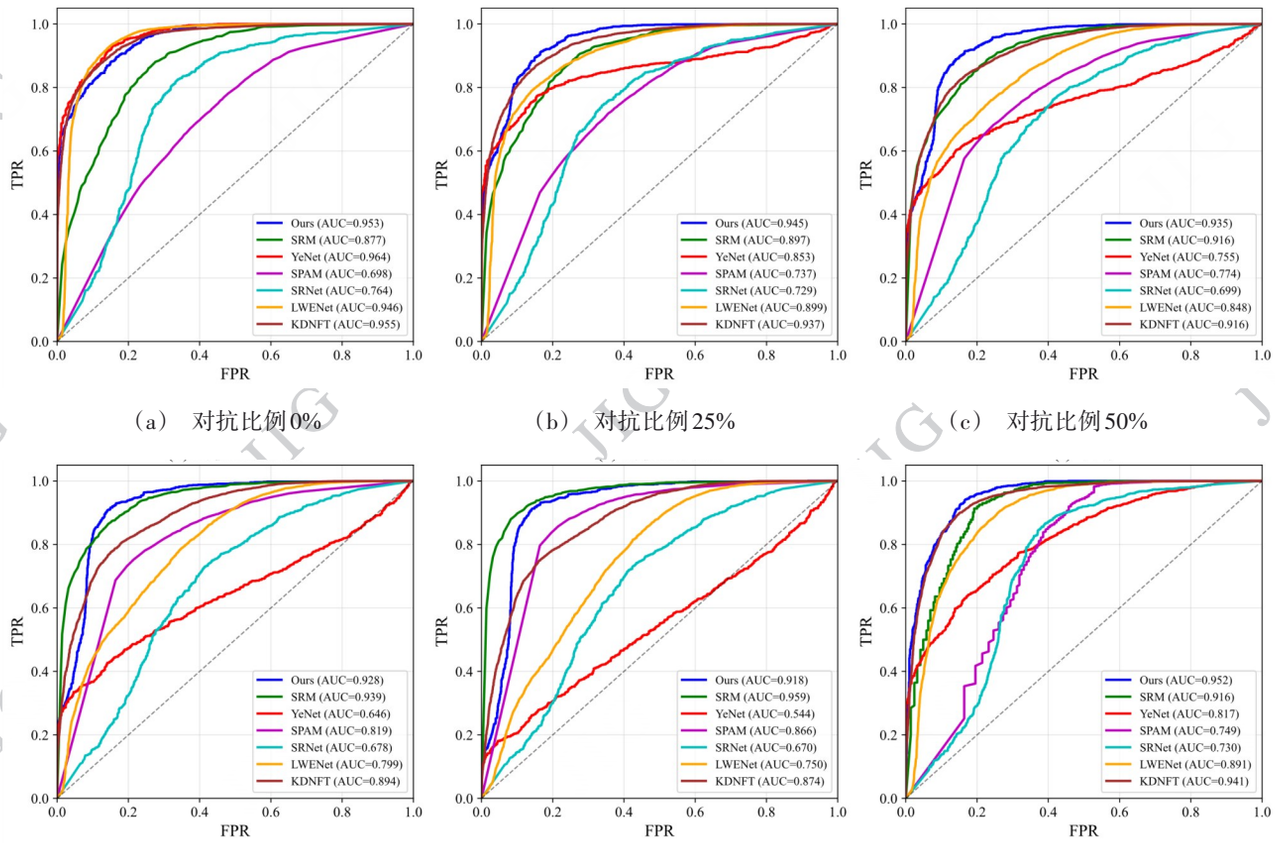
综合而言S-UNIWARD场景下不同对抗比例与三种嵌入率(0.2bpp、0.4bpp和0.6bpp)以及HILL场景下的实验结果进行整体平均分析，融合模型$p^{E0.5}$均值为0.1913。相比于SPAM与SRM两种传统隐写分析模型，分别降低了0.1595和0.0606，相比于Ye-Net、SRNet与LWENet三种深度学习隐写分析模型分别降低了0.1511、0.1950和0.1093，相比于KDNFT，降低了0.0590。在综合考虑不同隐写算法、不同嵌入率以及不同对抗强度的复杂实验环境下，所提出模型在整体错误率上显著优于现有传统方法与主流深度学习方法，在对抗隐写分析场景中实现了当前最优(SOTA)检测性能。同时，该优势并非来源于特定算法或单一嵌入率条件，而是在多场景、多强度对抗环境下保持稳定领先，体现出良好的泛化能力与鲁棒性。

2.4 对抗测试

在这一部分中，分别以SRNet、LWENet与KDNFT作为攻击目标构造对抗隐写图像，嵌入率设定为0.6bpp，隐写算法采用S-UNIWARD与ADV-EMB。由于SPAM与SRM属于传统统计型隐写分析方法，不具备端到端的梯度传递结构，因此无法基于梯度信息生成针对性的对抗隐写图像。对于KDNFT而言，其结构包含深度学习分支与传统特征分支两个通道，因而仅能针对其中的端到端深度学习部分构造对抗扰动。

对于融合模型而言，由于在集成学习模块与DANN分类器衔接处引入了不可导的偏离样本过滤机制，对于偏离样本完全使用SRM得到最终测试结果，该机制在训练与推理过程中截断了梯度的连续传播路径，使得基于梯度的白盒对抗攻击难以直接实施。因此无法直接依赖梯度传递制作对抗隐写图像。考虑到融合模型的性能在一定程度上依赖基学习器输出表现，本文采用针对Ye-Net构造的对抗隐写图像进行检测评估。在高嵌入率条件下(0.6bpp)，Ye-Net与SRM均会受到不同程度的干扰，从而间接影响融合模型的整体判别结果，实现对融合框架的间接对抗测试。最终实验结果如表3所示。

随着对抗比例的逐步升高，包括融合模型在内的所有方法均呈现出不同程度的性能退化趋势，且



（a） 对抗比例0%　（b） 对抗比例25%　（c） 对抗比例50%

（d） 对抗比例75%　（e） 对抗比例100%　（f） 平均

((a) Adversarial Ratio 0%; (b) Adversarial Ratio 25%; (c) Adversarial Ratio 50%; (d) Adversarial Ratio 75%; (e) Adversarial Ratio 100%; (f) Average)

图10　不同对抗比例下各模型的ROC曲线和AUC值

Fig. 10　ROC Curves and AUC Values of Each Model Under Different Adversarial Ratios

在高对抗比例下更为显著。具体而言，SRNet与LWENet在无对抗（对抗比例0%）向着满对抗（对抗比例100%）的变化过程中，$p^{E0.5}$值分别上升了0.4347与0.4870，表现出对对抗扰动较高的敏感性。相比之下，KDNFT与所提出的融合模型$p^{E0.5}$值的波动幅度相对较小，分别上升0.1517与0.1635，整体变化较为平缓。从对对抗隐写图像的鲁棒性的角度来看，KDNFT在性能稳定性方面略占优势，其指标增幅最小；然而，从全局检测性能及平均误差水平来看，融合模型在绝大多数对抗比例条件下仍保持更优或次优表现，整体性能优于KDNFT。由此可见，所提出方法在保证较强鲁棒性的同时，兼顾了检测精度与稳定性，在综合性能层面具有更明显优势。

2.5　复杂度分析

在所提出的模型框架中，引入集成学习能显著提升模型的检测性能，但直接堆叠多个Ye-Net和SRM基学习器会带来难以承受的计算负担。一方面，Ye-Net分支的计算成本主要源于深度卷积网络执行的大量卷积运算。若采用朴素的集成方法独立

表3　对抗测试下不同模型$p^{E0.5}$值对比

Table 3　Comparison of $p^{E0.5}$ Values of Different Models under Adversarial Testing

对抗比例	SRNet	LWENet	KDNFT	Ours(Ye-Net)
0%	0.159 8	0.110 3	0.131 9	**0.089 5**
25%	0.267 9	0.231 9	0.171 2	**0.129 0**
50%	0.374 2	0.354 1	0.206 2	**0.176 0**
75%	0.481 7	0.475 8	0.246 5	**0.210 0**
100%	0.591 5	0.597 3	0.283 6	**0.253 0**
平均	0.375 0	0.353 9	0.207 9	**0.171 5**

训练多个Ye-Net网络，总训练时间将成倍增长，这使得模型在时间上不具备可行性。另一方面，SRM分支的瓶颈在于高维人工特征的构建。该过程需对全图进行多阶残差滤波与贡献矩阵统计。在传统的集成框架下，训练多个SRM模型意味着需要对同一张图像重复执行多次相同的特征提取操作，这种计算上的冗余极大地限制了模型的整体运行效率。

为了量化模型各组件的实际计算负载，本文在配置为RTX 2080Ti GPU的计算平台上对关键模块的时间消耗进行了统计，统计范围包括：

1）使用完整训练集执行单次Ye-Net网络训练和单次SRM模型构建（含特征提取和集成FLD分类器训练）的时间。

2）单个Ye-Net模型和SRM模型对单张输入图像完成推理所需的平均时间。

3）排除特征提取与集成学习部分的耗时后，独立运行DANN分类器（含MLP过滤器）的训练推理时间。

具体统计结果如表4所示：

表4　时间消耗统计

Table 4　Time Consumption Statistics

	训练时间	推理时间
Ye-Net	28.82h	2.967s
SRM	30.17h	24.42ms
DANN分类器	185.8s	1.478ms

从表4的统计数据可以看出，Ye-Net分支与SRM分支的训练耗时占据了绝大部分计算资源，而引入MLP过滤器和DANN分类器仅产生了极微小的边际成本。因此，本文针对Ye-Net与SRM两大分支的训练过程进行优化，使得提出的集成框架不会因引入多个基学习器而导致上述时间成本的线性增长。

针对Ye-Net分支，本文引入快照集成策略（Huang等，2017），即从单次训练过程的不同收敛阶段获取基学习器，从而使总训练成本并未因集成规模的扩大而显著增加。在具体实现中，共执行了一次独立的Ye-Net训练过程，根据验证集上的$P_E^{0.5}$值对保存的模型快照进行排序，并选取性能最佳的十个快照作为基学习器，参与模型后续的训练和推理流程。

针对SRM分支，本文采用特征复用机制。对于任意训练图像，其高维残差特征仅需提取一次并保存。后续的FLD集成训练基于对该特征向量的随机子空间采样与闭式矩阵求解，其计算过程不涉及迭代优化。本文在提取完成的特征空间中训练了多组FLD集成分类器，每组分类器的训练训练时间平均为25.22s，远低于特征提取所消耗的时间。因此，本文不再对FLD集成分类器的训练过程进行优化，以保留最佳效果。

此外，在集成学习阶段，Ye-Net分支与SRM分支完全独立，这为Ye-Net分支与SRM分支并行计算提供了支持。通过上述优化方案，整个模型的训练时长可以被控制在31h以内，这一时间开销与单独训练一个标准的Ye-Net网络或SRM模型基本持平，并未引入显著的时间负担。

在模型的推理阶段，计算开销分布呈现出显著的不均衡性，由于深度卷积网络需在高分辨率特征图上执行密集的前向传播计算，Ye-Net分支处理单张图像的耗时远高于SRM分支与DANN分类器，构成了在线检测的主要计算瓶颈。尽管采用集成策略使得Ye-Net分支的总推理时间随基学习器数量呈线性增加，集成10个YeNet模型最终对单张图像的推理时间约为30s，但在隐写分析任务中，对于检测精度的要求往往高于实时性要求，且鉴于各基学习器相互独立，工程部署时通过多GPU并行推理相对容易，因此，该推理开销仍在实际工程应用的可接受范围内。

2.6　组件效果测试

本节围绕融合框架中的关键组件开展组件效果测试：通过有针对性地移除或替换单个模块，评估其对整体检测性能的影响。所有测试均采用与主实验一致的数据划分与评测规定，并在不同对抗比例条件下重复进行。为保证各组结果具有可比性，除被调整的组件外，其余数据预处理流程、训练策略与评测指标均保持不变。具体设置如下：

实验1：MLP过滤器移除。移除偏离样本门控机制，即不再使用MLP过滤器对目标域样本进行偏离判定，使用全部目标域样本的Ye-Net预测概率矩阵训练DANN网络，在推理阶段所有样本均需要经过DANN网络得到DANN预测概率p_D和SRM投票结果p_s，且权重平均分配。

实验2:SRM分支移除。推理阶段移除SRM分支,不再区分偏离样本,只使用DANN网络输出作为最终判别结果,以评估SRM分支及融合机制对鲁棒性的提升贡献。

实验3:概率特征归一化移除。通过集成学习得到的预测概率矩阵不再进行归一化校准处理,而是直接将各基学习器的原始输出作为后续训练与推理的输入特征,用于评估概率特征归一化步骤对融合与跨域学习的影响。

实验4:归一化强度因子固定。保留Sigmoid归一化步骤,但归一化阶段直接使用标准Sigmoid函数进行归一化,不再针对Ye-Net与SRM两路特征分别设置不同的归一化强度因子。

表5汇总了完整模型与上述组件调整设置在相同评测标准下的定量结果,用以衡量各组件对最终性能的边际贡献,指标仍使用$P^{E0.5}$。整体来看,完整模型在多数对抗比例下均取得最优的表现,且随着对抗比例上升性能保持稳定,平均$P^{E0.5}$值为0.1271,显著优于四组组件移除实验的结果,证明各组件对整体性能均具有正向贡献。

表5 组件效果测试

Table 5 Component Effectiveness Evaluation

对抗比例	完整模型	实验1	实验2	实验3	实验4
0%	0.135 5	0.124 5	**0.109 5**	0.202 0	0.183 5
25%	**0.127 0**	0.172 5	0.192 5	0.184 5	0.171 5
50%	**0.125 5**	0.219 5	0.273 5	0.174 5	0.164 0
75%	**0.118 0**	0.267 5	0.360 0	0.157 0	0.147 5
100%	**0.117 0**	0.314 5	0.442 5	0.144 5	0.137 0
平均	**0.124 6**	0.226 3	0.275 6	0.172 5	0.160 7

MLP过滤器的作用主要体现在高对抗比例下的稳定性。随着对抗隐写图像数量的增加,偏离样本数量上升,显著干扰域对抗对齐过程。这体现在移除MLP过滤器后,随着对抗比例的增加,$P^{E0.5}$值由0.1215提升至0.3320,呈现出明显的上升趋势。通过过滤偏移过大的对抗隐写图像能有效缓解负迁移,使域适应在高攻击强度下仍能稳定工作。

SRM分支对强对抗环境至关重要,当移除SRM分支,仅依赖DANN进行判别时,误差率随对抗比例上升更为迅速,尤其在100%对抗比例下,移除SRM分支所导致的误差率约为完整模型的3.73倍。该结果说明在偏移较强的对抗隐写图像上,单纯依赖域适应得到的特征仍不足以覆盖全部攻击分布,尤其是针对在训练过程中被过滤掉的偏离样本。SRM分支提供了对对抗扰动更稳健的互补判别信息,是提升强对抗场景鲁棒性的关键来源。

概率归一化与归一化强度因子主要影响基学习器输出尺度对齐和整体融合的可用性。移除概率归一化会在各对抗比例下造成一致的性能下降,$P^{E0.5}$值平均提升0.0456。而直接使用标准Sigmoid函数进行归一化也会带来一定程度上的性能退化,$P^{E0.5}$值平均提升0.0325。因此,归一化强度因子能够缓解异构检测器输出分布不一致带来的尺度偏差,使融合输入更稳定,提升模型性能上限。

3 结 论

针对深度学习隐写分析模型在对抗隐写攻击下性能衰减的关键问题,本文通过系统实验验证了深度学习模型在面对对抗隐写时表现出的脆弱性,并提出了一套面向未知对抗比例、无真值目标域的鲁棒隐写分析模型。该模型以深度学习隐写分析模型Ye-Net和传统空域隐写分析模型SRM为基学习算法,采用集成学习框架进行有效融合,在此基础上,以DANN域对抗网络作为核心分类器进行跨域特征对齐,同时在DANN网络前引入基于MLP的偏离样本识别模块,对目标域中偏移过大、易在无监督域对抗训练过程中错误对齐至载体图像区域的样本进行过滤,以抑制负迁移;推理阶段结合偏离判定结果进行加权融合,使模型能够在不依赖目标域真值、也无需预先获知对抗强度的条件下,实现对对抗隐写图像的自适应鲁棒检测。基于充分的实验验证,我们得到以下结论:

(1)深度学习模型与传统模型在对抗环境下呈现出显著的性能互补性。实验数据表明,在无对抗的纯净环境下,深度学习模型LWENet和Ye-Net的检测误差$P^{E0.5}$均显著优于传统模型SRM,展现了其在特征拟合上的优势。然而,面对基于梯度的对抗攻击,深度模型的判别能力急剧退化,而传统模型的误差却反向降低,证明其基于残差统计的手工特征对对抗扰动引入的伪影具有较强的鲁棒性。

(2)融合隐写分析模型显著提升了模型针对对

抗隐写图像的检测性能。从平均性能来看，所提融合隐写分析模型的检测误差$P_{E0.5}$仅为0.1913，与传统隐写分析方法相比，该模型较SPAM与SRM的错误率分别下降了15.95%和6.06%；与深度学习模型相比，相较Ye-Net、SRNet和LWENet的错误率平均下降了15.11%、19.50%和10.93%，相较于针对对抗隐写的鲁棒性增强方法KDNFT错误率平均下降5.90%，在对抗隐写场景下达成当前SOTA的隐写分析性能。此外，在对抗测试下，融合隐写分析模型的平均检测误差仅为0.1715，取得了最佳检测性能。

（3）组件效果测试进一步说明，偏离样本过滤与DANN和SRM双分支融合机制是提升强对抗条件下鲁棒性的关键因素。在移除MLP过滤器或SRM分支后，模型$P_{E0.5}$值整体提升了0.0992和0.1566，尤其是在面对对抗隐写图像时，模型$P_{E0.5}$值提升了0.2100和0.3330，上述结果验证了偏离过滤和双分支融合能够有效提升模型在强对抗环境下的域对抗训练效果。

本研究为提升隐写分析模型在对抗环境下的鲁棒性提供了新的思路和方法，所提出的融合框架在安全性和实用性方面均展现出良好的应用前景。未来可以进一步优化多模型融合权重策略，探索更高效的跨域对齐网络结构；同时拓展至复杂载体图像与多样对抗攻击类型，提升模型的泛化适配能力。

参考文献（References）

Anderson R J and Petitcolas F A P. 1998. On the limits of steganography. IEEE Journal on Selected Areas in Communications, 16(4): 474-481 [DOI: 10.1109/49.668971]

Bas P, Filler T and Pevný T. 2011. "Break Our Steganographic System": The ins and outs of organizing BOSS// Information Hiding (Lecture Notes in Computer Science, vol. 6958). Berlin: Springer: 59-70 [DOI: 10.1007/978-3-642-24178-9_5]

Boroumand M, Chen M and Fridrich J. 2019. Deep residual network for steganalysis of digital images. IEEE Transactions on Information Forensics and Security, 14(5): 1181-1193 [DOI: 10.1109/TIFS.2018.2871749]

Chen J F, Fu Z J, Zhang W M, Cheng X and Sun X M. 2021. A survey of image steganalysis based on deep learning. Journal of Software, 32(2): 551-578 (陈君夫，付章杰，张卫明，程旭，孙星明. 2021. 基于深度学习的图像隐写分析综述. 软件学报，32(2)：551-578) [DOI: 10.13328/j.cnki.jos.006135]

Fridrich J and Kodovský J. 2012. Rich models for steganalysis of digital images. IEEE Transactions on Information Forensics and Security, 7(3): 868-882 [DOI: 10.1109/TIFS.2012.2190402]

Ganin Y, Ustinova E, Ajakan H, Germain P, Larochelle H, Laviolette F, et al. 2016. Domain-adversarial training of neural networks. Journal of Machine Learning Research, 17(59): 1-35

Goodfellow I J, Shlens J and Szegedy C. 2014. Explaining and harnessing adversarial examples[EB/OL].[2025-12-26]. https://arxiv.org/abs/1412.6572

Holub V and Fridrich J. 2013. Digital image steganography using universal distortion//Proceedings of the 1st ACM Workshop on Information Hiding and Multimedia Security. New York: ACM: 59-68 [DOI: 10.1145/2482513.2482514]

Hu M Z and Wang H X. 2023. Image steganalysis against adversarial steganography by combining confidence and pixel artifacts. IEEE Signal Processing Letters, 30: 987-991 [DOI: 10.1109/LSP.2023.3300792]

Huang G, Li Y, Pleiss G, et al. 2017. Snapshot ensembles: Train 1, get M for free[EB/OL].[2025-01-20]. https://arxiv.org/abs/1704.00109

IEEE Signal Processing Society. 2025. Break Our Watermarking System - 2nd Ed. (BOWS-2)[EB/OL].[2025-12-23]. https://signalprocessingsociety. org/publications-resources/data-challenges/break-our-watermarking-system-2nd-ed

Jawad T A, Mohasefi J B, Abdelghany M S R. Adversarial-robust steganalysis system leveraging adversarial training and EfficientNet[J]. TELKOMNIKA Telecommunication Computing Electronics and Control, 2025, 23(2): 393-401. DOI: 10.12928/TELKOMNIKA.v23i2.26614.

Li C and Wang B. Fisher linear discriminant analysis[J]. CCIS Northeastern University, 2014, 6.

Li B, Wang M, Huang J, Li X. 2014. A new cost function for spatial image steganography. IEEE International Workshop on Information Forensics and Security (WIFS): 1 - 6 [DOI: 10.1109/WIFS.2014.7084304]

Lin K, Li W, Barn M, et al. 2024. Constructing an intrinsically robust steganalyzer via learning neighboring feature relationships and self-adversarial adjustment. IEEE Transactions on Information Forensics and Security, 19: 9390-9405 [DOI: 10.1109/TIFS.2024.3470651]

Liu M, Luo W, Zheng P, et al. 2021. A new adversarial embedding method for enhancing image steganography. IEEE Transactions on Information Forensics and Security, 16: 4621-4634 [DOI: 10.1109/TIFS.2021.3111748]

Long Linghui, Wang Zichi, Zhang Xinpeng. 2026. Overview of neural network model steganography. Journal of Image and Graphics, 31(1):0045-0061 (龙玲慧，王子驰，张新鹏. 2026. 神经网络模型隐写研究进展. 中国图象图形学报，31(1)：0045-0061) DOI: 10.11834/jig.250267.DOI: 10.11834/jig.250267.

Ma Bin, Li Kun, Xu Jian, Wang Chunpeng, Li Jian, Zhang Liwei.

2024. High-security image steganography with the combination of multiple competition and channel attention. Journal of Image and Graphics, 29(02):0355-0368 (马宾, 李坤, 徐健, 王春鹏, 李健, 张立伟. 2024. 联合多重对抗与通道注意力的高安全性图像隐写. 中国图象图形学报, 29(02):0355-0368).DOI: 10.11834/jig.230134.

Ma S, Zhao X and Liu Y. 2019. Adaptive spatial steganography based on adversarial examples. Multimedia Tools and Applications, 78: 32503-32522 [DOI: 10.1007/s11042-019-07994-3]

Petitcolas F. 1883. La cryptographie militaire. Journal des Sciences Militaires, 9: 161-191

Pevný T, Bas P and Fridrich J. 2010. Steganalysis by subtractive pixel adjacency matrix. IEEE Transactions on Information Forensics and Security, 5(2): 215-224 [DOI: 10.1109/TIFS.2010.2045842]

Qin C, Zhang W, Dong X, et al. 2021. Adversarial steganography based on sparse cover enhancement. Journal of Visual Communication and Image Representation, 80: 103325 [DOI: 10.1016/j. jvcir. 2021.103325]

Qin C, Zhao N, Zhang W, et al. 2022. Patch steganalysis: A sampling based defense against adversarial steganography//Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE: 3079-3083 [DOI: 10.1109/ICASSP43922.2022.9747638]

Sharma V, Mir R and Rout R. 2023. Towards secured image steganography based on content-adaptive adversarial perturbation. Computers & Electrical Engineering, 109: 108757 [DOI: 10.1016/j.compeleceng.2023.108757]

Suykens J A K and Vandewalle J. 1999. Least squares support vector machine classifiers. Neural Processing Letters, 9: 293-300 [DOI: 10.1023/A:1018628609742]

Tang W, Li B, Tan S, et al. 2019. CNN-based adversarial embedding for image steganography. IEEE Transactions on Information Forensics and Security, 14(8): 2074-2087 [DOI: 10.1109/TIFS.2019.2891237]

Weng S, Chen M, Yu L and Sun S. 2022. Lightweight and effective deep image steganalysis network. IEEE Signal Processing Letters, 29: 1888 - 1892 [DOI: 10.1109/LSP.2022.3201727]

Xu G, Wu H Z and Shi Y Q. 2016. Structural design of convolutional neural networks for steganalysis. IEEE Signal Processing Letters, 23(5): 708-712 [DOI: 10.1109/LSP.2016.2548421]

Ye J, Ni J and Yi Y. 2017. Deep learning hierarchical representations for image steganalysis. IEEE Transactions on Information Forensics and Security, 12(11): 2545-2557 [DOI: 10.1109/TIFS. 2017.271094

Zhang Y, Zhang W, Chen K, Liu J, Liu Y and Yu N. 2018. Adversarial examples against deep neural network based steganalysis//Proceedings of the ACM Workshop on Information Hiding and Multimedia Security. New York: ACM: 67-72 [DOI: 10.1145/3206004.3206012]

作者简介

田华伟,男,博士,教授,主要研究方向为信息隐藏与多媒体取证、网络安全、公安情报。E-mail:hwtian@ppsuc.edu.cn