NeuroShield：一种用于对抗鲁棒性的神经符号框架

Ali Shafiee Sarvestani[1], **Jason Schmidt**[2], **和 Arman Roohi**[1*]

[1]伊利诺伊大学芝加哥分校，电气与计算机工程系，芝加哥，伊利诺伊州 60607，美国[2]伊利诺伊大学芝加哥分校，计算机科学系，芝加哥，伊利诺伊州 60607，美国[*]aroohi@uic.edu

arXiv:2601.13162v1 [cs.LG] 19 Jan 2026

摘要

对抗脆弱性和缺乏可解释性是深度神经网络面临的关键局限，尤其是在自动驾驶等安全敏感场景中。我们提出 **NeuroShield**，一种神经符号框架，将符号规则监督整合到神经网络中，以同时提升对抗鲁棒性和可解释性。我们将领域知识编码为关于形状、颜色等外观属性的逻辑约束，并通过在训练期间施加语义损失和符号逻辑损失来强制执行这些约束。利用 GTSRB 数据集，我们在标准ℓ_∞扰动预算 $\varepsilon = 8/255$下评估模型抵御快速梯度符号法和投影梯度下降攻击的鲁棒性。与干净样本训练相比，标准对抗训练仅带来了有限的鲁棒性提升（∼10 个百分点）。相比之下，我们的 FGSM-神经符号和 PGD-神经符号模型分别较其对应的对抗训练基线将对抗准确率提升了 18.1% 和 17.35%，取得的鲁棒性增益约为标准对抗训练的三倍；两者均以相同的干净样本训练基线为参照进行测量，且不会降低干净样本准确率。与 LNL-MoEx 等基于 Transformer 的防御方法相比，后者需要复杂的架构和大规模数据增强，而我们的 PGD-神经符号变体仅使用经过 10 轮次训练的 ResNet18 骨干网络，便达到了相当甚至更高的鲁棒性。这些结果表明，符号推理为构建稳健且可解释的人工智能提供了一条有效路径。

引言

深度学习模型的黑箱特性使其容易受到对抗攻击和各种失效模式的影响。模型缺乏透明度，限制了人类可解释性和人工监督。传统的人工监督主要局限于数据集准备和预处理，这对于需要可信度和可解释性的安全敏感任务而言远远不够。

神经符号人工智能（NSAI）通过为深度神经网络（DNN）增添符号推理，提升可解释性、可信度和鲁棒性[1]。深度神经网络可以从大型数据集中提取抽象特征，并学习复杂的决策边界[2]。然而，其内部机制依然不透明，用户只能信任模型的预测，却无法深入了解决策过程。这种不透明性可能导致模型在现实应用中造成有害错误。神经符号人工智能通过将特定领域的符号规则纳入神经模型来应对这一问题。这些符号组件充当专家监督器，用于验证模型的输出是否符合已知的逻辑关系。例如，在交通标志分类中，符号规则可以要求“STOP”标志必须是红色的八边形。如果模型在未检测到这些属性的情况下预测了某个类别，符号模块便可标记这一不一致，如图1所示。根据集成策略的不同，符号组件可以充当事后验证器、并行逻辑引擎，也可以作为影响训练动态的嵌入式结构。有些方法将逻辑融入损失函数，另一些则利用符号规则对输出进行筛选或重新排序[3,4]。神经符号人工智能系统可以依赖预先定义的规则（类似于监督学习），也可以通过数据驱动的归纳机制，在训练期间动态提取新规则。尽管表现出色，深度神经网络仍极易受到对抗威胁。这些威胁源于微小且蓄意设计的扰动，这些扰动通常肉眼难以察觉，目的是误导模型。这种脆弱性源于模型依赖表面的统计相关性，而非语义推理。因此，即使是训练有素的模型，也可能被与干净样本几乎完全相同的输入欺骗。对抗攻击可能发生在训练和推理阶段。有些攻击通过注入恶意样本破坏训练，另一些则针对推理阶段，诱使模型给出置信度高却错误的预测[5]。这些攻击既可以设计成范围广泛且随机的，也可以高度针对特定输出。无论哪种情况，它们都会利用模型缺乏推理能力和可解释性的弱点，暴露神经模型的脆弱性。

为此，人们开发了各种防御策略，以提高模型的韧性。这些策略通常包括：调整训练过程，让模型接触带噪声或经过扰动的输入；在测试时修改输入，以抑制对抗模式；或者改变模型架构，掩盖攻击者所依赖的梯度。一些方法试图通过形式化验证技术来保证鲁棒性，另一些则旨在降低模型对输入变化的敏感度[5]。然而，所有这些方法都仅在神经域中运作，且仍受训练数据的限制。这

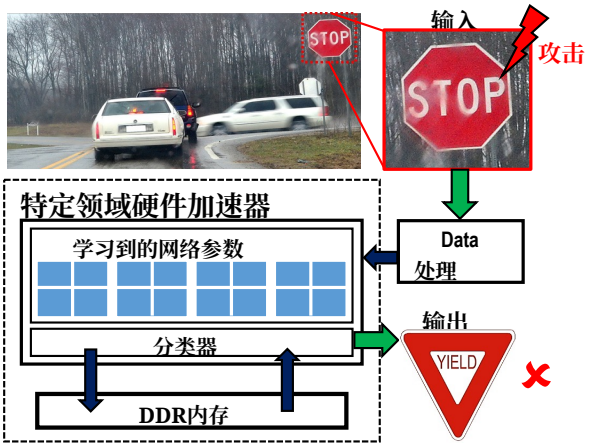

图1. 神经网络加速器中的软错误可能导致图像分类错误，使自动驾驶汽车可能在应急制动时加速而非突然刹车。（改编自[6]）。

一局限促使人们将符号推理融入学习。与纯统计模型不同，神经符号系统能够从逻辑和结构化知识的角度评估预测。遇到对抗性输入时，这类系统并非完全依赖已学习的模式，也可以依据代表领域知识的逻辑约束。这些约束有助于确保输出符合基本关系，例如已知交通标志的形状和颜色要求。即使扰动输入误导了神经组件，符号模块也会充当语义过滤器，识别其中的矛盾。这样一来，推理能力便提供了额外的防御层：通过基于可解释、可验证的规则而非纯统计特征得出预测，增强系统的鲁棒性。

本文提出了**NeuroShield**，这是一种神经符号人工智能框架，旨在通过将符号推理融入学习流程，提升神经网络的可解释性和鲁棒性。本文其余部分安排如下：第1节回顾神经符号学习和对抗鲁棒性方面的相关工作；第2节介绍我们的基于规则的监督框架；第3节和第4节详细介绍骨干网络修改和训练流程；第6节评估我们模型的鲁棒性、可解释性和效率。

1 相关工作

由于我们的方法将符号推理融入神经网络，以增强其对抗鲁棒性，因此我们将相关工作分为两部分：符号推理，回顾将逻辑嵌入神经学习的神经符号框架；以及对抗防御，综述抵御对抗攻击的技术。

1.1 符号推理

近年来，神经符号人工智能领域的进展探索了将符号推理与神经网络学习相结合，以提升可解释性、一致性和泛化能力。逻辑张量网络（LTN）[3] 通过一种名为 Real Logic 的可微逻辑系统，将逻辑符号（常量、函数、谓词）嵌入实值张量，从而结合一阶逻辑与深度学习。逻辑张量网络支持模糊真值、逻辑运算和量词，并已应用于分类、回归和半监督任务。神经逻辑机（NLM）[7] 通过一种用于关系推理的可微架构扩展了这一理念，利用权重共享和池化来学习类逻辑规则，并在传递推理和程序归纳等任务之间实现泛化。

神经符号前向推理（NSFR）[8] 通过从图像中提取以对象为中心的表示，并借助谓词分类器将其映射为逻辑原子，将符号逻辑与感知相结合。这些原子由一个可微前向链接引擎处理，该引擎基于模糊真值执行多步推理，从而在 Kandinsky Patterns 和 CLEVR-Hans 等视觉推理任务中取得可解释的结果。不同于仅依赖潜在特征的方法，NSFR 强调以感知为基础的规则驱动行为。语义损失[9] 通过定义作用于命题逻辑公式的可微损失，提供了另一种嵌入符号约束的方法。它会惩罚对违反逻辑的输出赋予高置信度的行为，并已在半监督分类以及路径生成、偏好排序等结构化预测任务中取得良好效果。

最终，神经符号验证[10] 通过将神经网络与基于逻辑的断言检查相结合，把符号推理应用于模型验证。规范网络定义语义条件，而验证规则则编码在

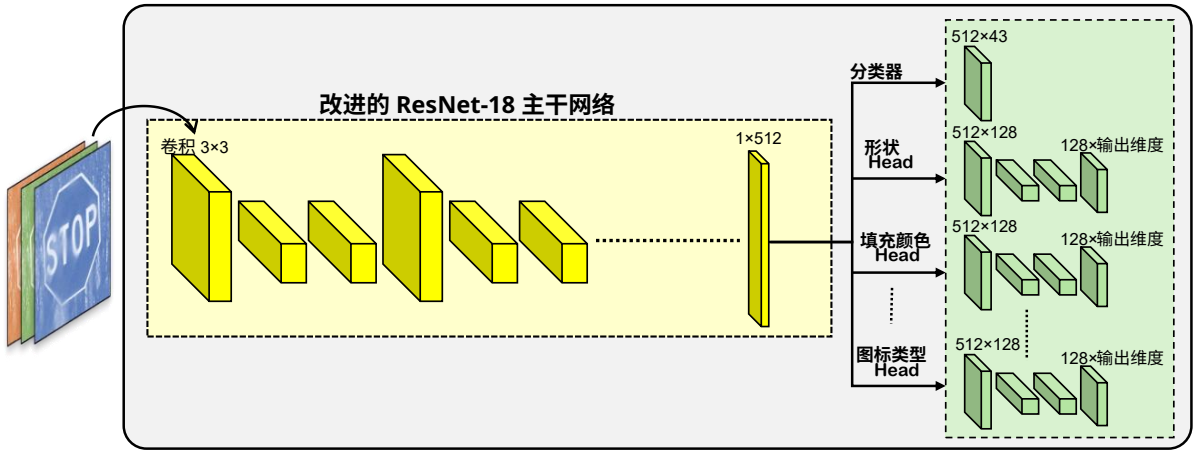

图2。 我们框架中使用的改进版 ResNet18 架构。该网络通过将初始 7×7 卷积替换为 3×3 卷积核并移除最大池化层，在早期层保留较高的空间分辨率。移除最终的全连接层后，所得的 512 维特征向量由交通标志分类器和符号属性头共享。

形式化断言语言中，并使用 SMT（理论可满足性）求解器进行检查，从而实现对安全关键系统的语义属性和多网络属性的验证。

1.2 对抗鲁棒性

本节回顾旨在提升对抗攻击鲁棒性的防御策略，包括对抗训练、输入变换、对抗样本检测和认证防御。近期研究表明，对抗训练仍是最有效的防御机制之一。文献 [11], 提出的快速梯度符号法（FGSM）通过单步梯度符号运算生成对抗样本，并在 MNIST 上显著提升了鲁棒性：在保持较高干净准确率的同时，对抗准确率提高了 60 多个百分点。然而，FGSM 训练可能导致梯度掩蔽，且无法抵御更强的迭代攻击。为克服这些局限，文献 [12], 将对抗训练形式化为鲁棒优化问题，并提出了投影梯度下降（PGD），将其作为一种通用的一阶对手。基于 PGD 的对抗训练显著提升了鲁棒性，在 MNIST 上针对 PGD攻击的精确率最高可达 89%，并成为对抗防御事实上的基准方法。尽管如此，PGD 训练的计算开销很大，且可能对训练时所用的特定攻击过拟合，从而降低其应对未见攻击或自适应威胁的效果。

除对抗训练外，输入变换也是一种实用的防御措施，旨在分类前去除扰动。常见技术包括 JPEG压缩、位深度降低、像素丢弃和总变分最小化。这些方法与模型无关且易于应用，但往往会降低干净准确率，且仍容易受到自适应攻击。研究表明，许多基于输入的防御措施依赖于梯度混淆，这可能带来虚假的安全感，也可能被精心设计的攻击绕过[13]。另一种防御思路是对抗样本检测，即利用辅助模型或统计检验，通过检查特征激活、梯度或输出分布来识别对抗性输入。尽管检测方法有时有效，但往往无法推广到不同的数据集或攻击，并且在白盒设置下尤其薄弱，因为检测器可以被直接攻击[14]。最后，认证防御旨在通过设计模型，为模型在给定范数球内抵抗扰动提供形式化鲁棒性保证。随机平滑、Lipschitz正则化和凸松弛等技术都属于这一类别。尽管这些方法能够提供可证明的鲁棒性，但计算开销很大，也难以扩展到大型模型和复杂数据集[15]。

2 规则表示与符号监督

在本文中，我们使用符号规则表示先验知识，这些规则将每个交通标志类别与其预期属性相关联，包括形状、填充颜色和边框颜色、图标或数字类型以及类别。这些规则构成一个轻量级知识库，使模型能够理解预测在语义上意味着什么。例如，停止标志应为红色、八边形，包含带有中央“STOP”图标的白色文本，具有白色边框，并属于优先通行类别。这些关联通过按类别字典静态定义（例如 class_shape_map、class_color_parts），以编码每种符号属性的预期值。每条规则都被转换为如下形式的软蕴含：

$$\text{if class} = c \Rightarrow \text{shape} = s,\ \text{color(fill)} = f,\ \text{category} = k,\ \ldots \tag{1}$$

我们引入两种形式的符号正则化，以提升可解释性和鲁棒性。首先，受语义损失的启发，[9]，我们定义了一种简化变体，用以鼓励预测类别分布与一组符号等价类别保持一致。该损失不会强制模型始终预测某个确切标签，而是允许模型在所有符号有效类别（例如，所有限速标志）之间分配置信度。这使模型的错误更具可解释性。给定真实标签 y，我们找出所有共享相同符号属性的类别，例如形状、颜色、图标类型和类别，并构造一个软目标分布 q：

$$q_i = \begin{cases} \frac{1}{|\mathscr{S}(y)|} & \text{if } i \in \mathscr{S}(y) \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

这里，$\mathscr{S}(y)$ 表示与标签 y 符号等价的类别集合，而 $q \in \mathbb{R}^{43}$ 定义了类别上的软目标分布。

语义损失通过计算该符号目标q 与预测类别分布 p_{cls}之间的 Kullback–Leibler散度得出：

$$\mathscr{L}_{\text{semantic}} = \text{KL}(q \parallel p_{\text{cls}}) \tag{3}$$

Kullback-Leibler（KL）散度用于衡量一个概率分布与另一个期望概率分布之间的差异。在此语境下，当模型对与真实标签在符号层面不一致的类别赋予较高置信度时，该散度会对模型进行惩罚。

其次，我们引入符号逻辑损失，以确保预测类别与其关联的符号组成部分保持一致。我们不采用硬性限制，而是使用受 Pylon框架[16]启发的软蕴含。对于每个组成部分，我们计算一个软真值：

$$T(A \Rightarrow B) = \begin{cases} \frac{p_B}{p_A + \varepsilon} & \text{if } p_A > p_B \\ 1 & \text{otherwise} \end{cases} \tag{4}$$

这里，p_A 表示对预测类别的置信度，而 p_B 表示对关联符号属性的置信度。$T(A \Rightarrow B) \in [0,1]$表示符号蕴含成立的程度。该公式确保模型在对某个类别作出高置信度预测时（即 p_A 较高），也必须对相应的符号属性赋予较高置信度（即 p_B）。如果 p_B 明显低于p_A，蕴含真值就会下降，表明存在违反，并导致损失增大。相反，如果p_A 较低，或者 $p_B \geq p_A$，则视为蕴含成立，不施加惩罚。公式（4）定义的这一软蕴含机制，贯彻了高置信度预测必须在语义上保持一致的原则。由此产生的符号逻辑损失使用对数函数对这些违反行为进行惩罚。

$$\mathscr{L}_{\text{logic}} = -\log(T(A \Rightarrow B) + \varepsilon) \tag{5}$$

在初步实验中，这种独立的逻辑一致性损失对鲁棒性的影响有限，主要是因为每个符号属性都是单独约束的。因此，我们引入了联合规则损失，将某个类别的全部符号属性视为一个整体结构化约束，对组合属性无法支持预测标签的预测结果进行惩罚。联合损失不再分别检查形状、颜色、类别和图标类型，而是评估整个属性配置在逻辑上是否与该类别一致。这样一来，它将基本逻辑损失从针对单个软蕴含的约束扩展为统一的多属性一致性约束。联合符号逻辑损失定义如下：

$$\begin{aligned} \mathscr{L}_{\text{joint}} &= -\log(T_{\text{joint}} + \varepsilon) \\ T_{\text{joint}} &= \min\left(\frac{\prod_j p_j^{w_j}}{p_{\text{cls}} + \varepsilon},\ 1\right) \end{aligned} \tag{6}$$

其中，p_j是符号属性j的预测概率，w_j是可调重要性权重。分子表示加权符号一致性，分母则确保相对于类别置信度进行归一化。为使逻辑损失更具适应性，我们在每个轮次根据模型的符号准确率更新公式（6）中的权重w_j。准确率较低的组成部分（例如形状、文本或图标）会获得更高的权重，从而促使模型在训练过程中更加专注于纠正这些不一致之处。

此外，通过这些自适应权重w_j，由于损失会自动对当前最薄弱的符号组件施加更大的压力，这会促使骨干网络组织其特征，使每个交通标志类别都占据

共享表示空间中一个独特且符合规则的区域，而不是依赖脆弱的低级线索。最终，我们使用总损失将符号正则化与标准监督学习相结合。

$$\mathscr{L}_{\text{total}} = \mathscr{L}_{\text{CE}} + \lambda_{\text{logic}} \cdot \mathscr{L}_{\text{logic}} + \lambda_{\text{semantic}} \cdot \mathscr{L}_{\text{semantic}} \tag{7}$$

其中， $\mathscr{L}_{\text{CE}}$ 是分类交叉熵， λ_{logic} 和 $\lambda_{\text{semantic}}$ 是控制逻辑项和语义一致性项影响的超参数。在我们的公式中，逻辑项L_{logic} 对应公式（6）中定义的联合符号损失。为便于说明，我们先在公式（5）中介绍基本的软蕴涵损失，然后将其扩展为公式（6）中的统一联合符号损失，这也是训练期间使用的版本。通过软约束和语义正则化整合符号规则，我们的框架提升了可解释性和对抗鲁棒性。

3 骨干网络修改与符号头部设计

为同时预测交通标志类别及其符号属性（如形状、颜色和图标类型），我们调整了标准 ResNet18 架构，将其作为语义分类和基于规则的监督共用的骨干网络。具体而言，为保留早期特征图中的空间细节，我们进行了两项结构调整：将初始 7×7 卷积替换为 3×3 卷积核，并移除最大池化层。这一调整减少了早期下采样，并保留了更高的空间分辨率，这对于识别细粒度的小型交通标志至关重要。我们还移除了最终的全连接（FC）分类层，并将其替换为恒等映射。因此，ResNet 从最终的全局平均池化层生成一个 512 维特征向量。该向量记为$\mathbf{f} \in \mathbb{R}^{512}$，作为所有下游头部的输入。这个共享特征向量$\mathbf{f}$ 会同时传入：

- 一个用于交通标志识别的标准分类头，以及
- 多个符号属性头，每个头都由一个轻量级多层感知机（MLP ）.

每个符号多层感知机头由以下部分组成：

- 一个用于降低维度（512 → 128）的线性投影层，
- 批归一化，
- ReLU 激活，以及
- 一个最终线性层，将输出映射到特定符号属性的输出空间（例如形状数量或颜色类别）。

形式上，每个符号头都可以表示为：

$$\text{多层感知机}_{\text{attr}}(\mathbf{f}) = W_2 \cdot \text{ReLU}(\text{BN}(W_1 \cdot \mathbf{f}))$$

其中$W_1 \in \mathbb{R}^{128\times512}$, $W_2 \in \mathbb{R}^{d\times128}$，且$d$ 是该属性的输出类别数（例如 5 种形状或 6 种颜色）。需要强调的是，我们不会手工设计符号特征，也不会将其作为输入。所有表征都直接从原始图像中学习得到。符号监督通过辅助输出及相应的基于逻辑的损失引入，同时模型仍保持完全端到端且可微。图2 展示了改进后的架构。共享的改进后 ResNet 骨干网络提取全局特征向量，随后由并行头用于类别预测和符号预测。

4 符号监督与对抗监督的训练策略

我们使用干净样本和对抗样本，在标准分类目标与符号规则监督相结合的指导下，对完整的德国交通标志识别基准（GTSRB）[17] 数据集训练模型。每个输入都经过修改后的 ResNet18 主干网络处理，并编码为一个 512 维特征向量，该向量由分类器头和多个符号头共享，用于预测形状、颜色组成部分、类别和图标类型等属性。为提升鲁棒性和可解释性，我们使用干净样本和对抗样本联合训练模型。对于每个批次，我们使用快速梯度符号法或投影梯度下降实时生成对抗样本，扰动强度为 $\varepsilon = 8/255$。这一数值带来了平衡的对抗威胁：扰动幅度足以对输入产生实质影响，同时在归一化图像尺度下仍无法被肉眼察觉。图3 展示了扰动幅度逐渐增大时，对抗扰动如何扭曲 ε 输入图像，同时在较低的扰动预算下仍基本无法被察觉。模型通过三种互补的损失函数进行监督：用于提升分类准确率的交叉熵损失、用于强制满足符号类别等价性的语义损失，以及用于确保遵循符号规则的逻辑一致性损失。这些损失均应用于干净样本和对抗样本，并采用公式（7）中定义的形式。

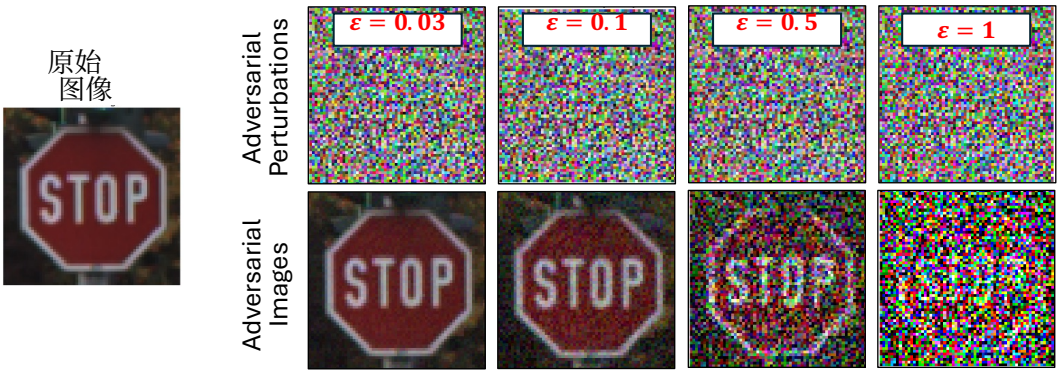

图 3. 原始交通标志图像，以及在不同扰动预算（$\varepsilon = 0.03$、0.1、0.5 和 1）下生成的对抗扰动和相应的对抗样本。扰动图展示了随着 ε 增大，像素级噪声如何加剧；生成的对抗图像则显示底层标志受到的破坏逐渐加重。

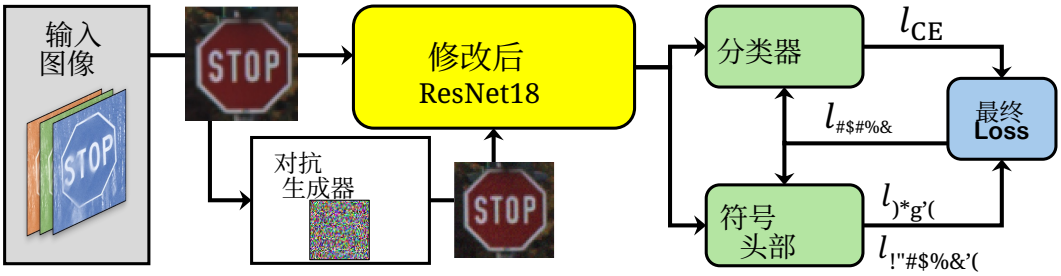

图 4. 训练策略示意图。每张输入图像都用于生成一个对抗样本，随后将干净样本和对抗样本输入共享的修改后的 ResNet18 主干网络，再经过分类器和符号头。最终损失结合了分类交叉熵、语义损失和符号逻辑损失，共同用于更新模型。

引入对抗训练后，联合规则损失尤为重要。标准的快速梯度符号法/投影梯度下降训练仅约束类别逻辑值在扰动下保持稳定；只要排名第一的标签仍然正确，骨干网络仍可通过脆弱的非语义特征传递梯度。在我们的框架中，每个对抗样本都会同时由分类器头和符号多层感知机头进行评估。如果某个对抗性输入保留了预测类别标签，却导致一个或多个符号属性（例如形状或类别）取值不一致，就会产生较大的联合逻辑损失。由于所有符号头共享同一个 512 维主干网络表征，来自L^{joint} 的梯度会促使骨干网络组织其特征，使模型通过语义一致的属性配置来实现鲁棒性，而不是以牺牲语义一致性为代价。换言之，联合规则损失本身并不会带来鲁棒性，但与对抗训练结合时，它会约束鲁棒性的实现方式，引导网络形成这样的决策区域：在攻击下，类别预测和符号属性都能共同保持稳定。

这种联合训练方案确保模型在自然和对抗条件下都能正确分类，同时保持其预测的符号一致性。与纯粹关注分类鲁棒性的传统对抗训练不同，我们的方法还会惩罚违反符号逻辑的行为，从而提升结构感知泛化。图4 展示了我们的训练流程。首先，将每张输入图像送入对抗生成器，以生成对应的对抗样本。随后，将干净样本和对抗样本都输入修改后的 ResNet18 主干网络，再由轻量级多层感知机头输出语义预测和符号预测。模型通过分类损失、语义损失和符号逻辑损失，在干净样本和对抗样本上接受监督；总损失则由这些分量的加权和计算得出。这种架构确保模型在标准条件和对抗条件下都具备鲁棒性和逻辑一致性。为了在最需要的地方进行符号正则化，我们会根据分类交叉熵和预测正确性，为每个批次计算自适应逻辑权重 λ_{logic}。这会加大对不确定或分类错误样本的逻辑约束压力，同时在模型置信度高且预测准确时降低这种压力。相比之下，语义损失权重$\lambda_{\text{semantic}}$ 保持固定，因为无论预测的置信度如何，其正则化效果都有益于整体表现。最后，虽然与标准对抗训练相比，符号推理会增加计算成本，但由此带来的可解释性和对抗鲁棒性提升足以证明这项开销是值得的，尤其是在交通标志识别等安全关键应用中。

表 1. 不同训练策略下干净准确率与对抗准确率的比较。

训练方法	干净准确率*	FGSM攻击准确率*	PGD 准确率*
未进行对抗训练	98.67%	36.7%	25.5%
仅使用快速梯度符号法训练	96.5%	45.0%	35.0%
仅进行PGD训练	4.5%	43.6%	38.65%
FGSM-干净样本训练	96.40%	46.70%	34.10%
PGD-干净样本训练	7.00%	44.3%	39.70%
神经符号（我们的方法）	97.90%	38.30%	19.90%
FGSM-神经符号（我们的方法）	98.20%	63.10%	42.68%
PGD-神经符号（我们的方法）	98.5%	65.75%	56.00%

*Clean Accuracy = 无攻击；**FGSM攻击准确率** = FGSM攻击下的准确率；**PGD准确率** =PGD攻击下的准确率

5 评估

我们在 GTSRB 数据集上评估了所提出的 **`NeuroShield`** 。该模型从头开始训练，仅训练 10 轮，且未进行任何预训练。这一刻意的限制有两个目的。首先，GTSRB 的视觉内容具有高度结构化的特点，因此模型能够快速收敛；训练时间过长可能导致模型对低级特征过拟合。其次，我们的目标是鼓励模型更多依赖符号推理和结构监督，而非记忆数据集特有的模式。尽管训练时间较短，我们的模型仍实现了较强的泛化能力，证明了符号正则化的有效性。我们从对抗鲁棒性和推理时间计算开销两方面评估了模型性能。首先，我们将其与标准对抗训练方法进行比较，结果表明，融入符号推理后，模型在快速梯度符号法和 PGD攻击下的鲁棒性显著提升，同时保持了干净准确率。我们还将该方法与近期一种基于 Transformer 的防御方法进行基准比较，结果显示，该方法准确率具有竞争力，甚至更优，同时训练成本大幅降低，可解释性也更好。最后，我们分析了神经符号模型的推理时计算开销，指出其在鲁棒性和效率之间实现了良好的权衡，并讨论了在资源受限环境中进一步降低计算复杂度的策略。

5.1 对比结果

5.1.1 我们的方法与标准对抗训练

我们在三种条件下评估了模型的鲁棒性：标准干净测试输入、使用 FGSM 生成的对抗样本，以及使用 PGD攻击生成的对抗样本。对于这两种攻击方法，我们采用的扰动强度为 $\varepsilon = 8/255$，它控制每个输入像素允许的最大变化量。选择此值是为了在显著扭曲输入的同时，使其在视觉上仍难以察觉，从而构成真实的对抗威胁。表 1 汇总了几种训练配置下的结果。该表第一列报告干净样本上的测试准确率，第二列和第三列分别显示模型在 FGSM 和 PGD攻击下的鲁棒性。我们的神经符号模型结合对抗训练（最后两行）在对抗样本上的准确率最高，同时在干净（非对抗）输入上仍保持超过 98% 的准确率，与表现最好的非对抗基线相当。

例如，与 FGSM 对抗训练（FGSM攻击准确率为 45.0%）相比，FGSM-神经符号的准确率达到 63.10%，提高了 +18.1%。同样，PGD-神经符号将 PGD 准确率从 38.65%（PGD 对抗训练）提升至 56.00%，提高了 +17.35%。这表明，符号规则监督显著增强了鲁棒性，效果优于标准对抗训练。我们的纯神经符号模型（未采用任何对抗训练）保持了较高的干净准确率（$\sim$ 97.90%），但在攻击下准确率有所下降，尤其是在 PGD 攻击下（$\sim$ 19.90%）。这验证了仅靠符号逻辑并不足够。然而，与对抗目标结合后，混合模型在攻击下的表现优于所有其他策略，展现了逻辑推理与对抗监督的互补性。

有趣的是，PGD-Clean 的干净准确率非常低（$\sim$ 7.00%），这可能是因为其中包含了强烈的对抗信号，扭曲了干净数据表示。尽管其 PGD 准确率尚可（$\sim$ 39.70%），但表现仍不如我们的混合模型。图 5（a）展示了我们的神经符号混合模型中干净样本和对抗样本的分布。对抗样本仍与对应的干净样本聚集得更紧密，这表明推理模块有助于保留语义结构，并在攻击下支持鲁棒且规则一致的预测。相比之下，图 5（b）展示了基线 ResNet 模型中的样本分布。对抗扰动破坏了内部特征空间的结构，导致聚类效果不佳，且干净样本与对抗样本之间存在显著重叠。

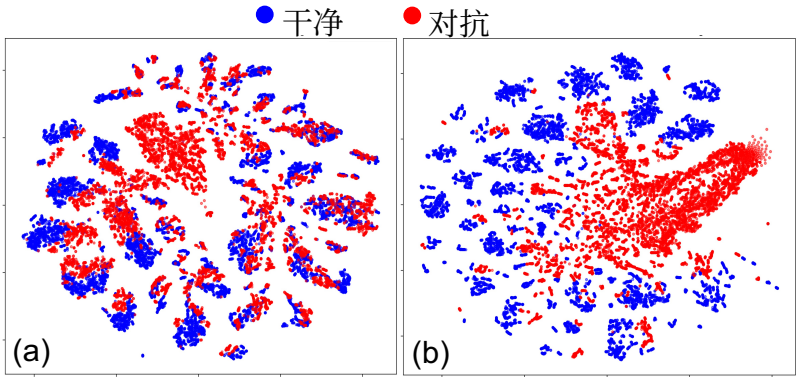

图 5. 使用 t-SNE 投影比较干净样本与对抗样本的分布。(a) 我们的神经符号模型能够更好地保留语义结构，使对抗样本与对应的干净样本仍然紧密对齐，凸显了基于规则的推理在维持鲁棒性方面的优势。(b) 基线 ResNet 模型的结构遭到破坏，干净样本与对抗样本之间存在显著重叠。

表 2. 不同方法在快速梯度符号法和 PGD攻击下的对抗准确率。我们将神经符号变体与基于 Transformer 的 LNL[18] 模型进行了比较，涵盖采用和未采用 MoEx 数据增强的情况。

方法	FGSM攻击准确率	PGD 准确率
LNL-Ti[18]	57.7%	37.9%
LNL-S[18]	64.5%	45.7%
LNL-MoEx-Ti[18]（数据增强）	71.3%	55.4%
LNL-MoEx-S[18]（数据增强）	77.8%	59.4%
FGSM-神经符号 ($\varepsilon = 0.03$)	63.10%	42.68%
FGSM-神经符号（$\varepsilon = 0.5$）	74.00%	25.00%
PGD-神经符号（$\varepsilon = 0.03$）	65.75%	56.00%

5.1.2 我们的方法与基于***Transformer***的模型对比

我们进一步将神经符号框架与*LNL-MoEx*[18]，进行比较。LNL-MoEx 是一种近期提出的基于 Transformer 的防御方法，它结合局部性归纳偏置与矩交换（MoEx）来增强对抗鲁棒性。LNL-MoEx 使用基于 ViT 的架构，在 GTSRB 上训练 100 轮次，在 CIFAR-10 上训练 150 轮次，并利用 MoEx 进行统计特征增强。值得注意的是，尽管许多基于 ViT 的防御方法依赖 ImageNet 预训练，LNL-MoEx 的作者强调，他们的方法是在 CIFAR-10 和 GTSRB 上从头训练的，未使用 ImageNet 预训练。

表 2 展示了在 FGSM 和 PGD攻击下对抗准确率的比较分析。在 $\varepsilon = 8/255 \sim 0.03$方面，我们的 PGD-神经符号模型取得了 56.00% 的 PGD 准确率，略高于 LNL-MoEx-Ti（55.4%），并接近表现最佳的 Transformer 基线 LNL-MoEx-S（59.4%），尽管无需数据增强且仅需进行 10 轮训练。在 $\varepsilon = 8/255$攻击下，我们的模型达到了 63.10% 的准确率，与 LNL-S（64.5%）相比具有竞争力，并显著优于 LNL-Ti（57.7%）。虽然 LNL-MoEx-S 取得了最高的 FGSM攻击准确率（77.8%），但它依赖 Transformer 主干网络和大规模数据增强[18]。然而，PGD攻击下的准确率更具意义，因为 PGD 在相同的 ε约束内执行多次迭代梯度步长，产生的扰动相比单步 FGSM攻击更有效且更接近对抗最优解。因此，PGD 鲁棒性被广泛视为更具挑战性且信息量更大的基准。

需要注意的是，LNL-MoEx 使用最高达 $\varepsilon = 1$ 的扰动强度来评估鲁棒性。这一扰动预算非常大，并不能反映交通标志识别等安全关键任务中的现实威胁水平。为在其扰动尺度内进行更公平的比较，我们训练了一个在 $\varepsilon = 0.5$ 下的 FGSM-神经符号变体，其 FGSM攻击准确率达到 74.00%；尽管我们的架构更简单、训练开销极小，但与 LNL-MoEx-S（77.8%）的差距已大幅缩小。然而，在 $\varepsilon = 0.5$ 下训练基于投影梯度下降的神经符号变体时，优化过程变得不稳定且未能收敛，这表明极大幅度的扰动可能需要额外的稳定化技术。尽管如此大的扰动预算在标准威胁模型中并不常见，而且如图 3 所示，较大的 ε 会严重破坏输入图像，我们仍使用 $\varepsilon = 0.5$ 仅用于与 LNL-MoEx 进行公平比较。

与 LNL-MoEx 相比，我们的模型具有以下几项优势：

表3. 在有无数据增强的情况下，各种基于Transformer的模型及我们的神经符号变体在快速梯度符号法和投影梯度下降攻击下的对抗准确率比较。

方法	FGSM攻击准确率	PGD 准确率
不使用数据增强		
PVT-Tiny[19]	10.8%	2.0%
TNT-T[20]	10.9%	3.9%
T2T-视觉变换器-t-10[21]	10.2%	0.4%
RVT-Ti[22]	31.5%	13.9%
Swin-T[23]	20.2%	7.8%
PVT-Small[19]	21.4%	2.4%
TNT-S[20]	21.2%	6.8%
T2T-视觉变换器-t-14[21]	21.7%	4.9%
RVT-S[22]	46.1%	25.3%
使用数据增强		
DeepAugment[24]	**65.1%**	42.2%
CutMix[25]	62.6%	40.1%
AugMix[26]	61.7%	39.8%
Puzzle-Mix[27]	**67.9%**	**44.2%**
RVT-Ti*[22]	38.9%	20.7%
RVT-S*[22]	51.2%	32.9%
神经符号（我们的模型）		
FGSM-神经符号（$\epsilon = 0.03$）	63.10%	42.68%
FGSM-神经符号 ($\epsilon = 0.5$)	74.00%	25.00%
PGD-神经符号 ($\epsilon = 0.03$)	65.75%	56.00%

- **可解释性**：LNL-MoEx 通过架构和统计技术增强鲁棒性，而我们的神经符号模型则通过强制确保语义规则一致性，同时提升鲁棒性和可解释性，并为每一次预测提供透明的推理依据。

- **轻量级正则化**：我们的符号监督模块与模型架构相互独立，可以轻松添加到任何神经网络骨干中，无需重新设计架构或调整融合策略。

- **高效训练**：我们的模型在 GTSRB 上仅使用 10 个训练轮次，便取得了有竞争力的对抗性能；而 LNL-MoEx 需要超过 100 个轮次，并采用复杂的 MoEx增强，才能达到类似结果。

为进一步说明我们结果的背景，表 3 比较了我们神经符号变体与多种基于Transformer的模型在有无数据增强情况下的对抗性能，相关结果引自先前研究[18]。具体而言，许多 Transformer 基线在未使用数据增强进行训练时，在快速梯度符号法和 PGD攻击下表现出较低的鲁棒性。即使采用 Puzzle-Mix 和 DeepAugment 等先进的数据增强策略，我们的 PGD-Neuro-Symbolic 模型仍实现了更优或相当的鲁棒性（PGD 下为 56.00%），同时保持轻量级架构和极低的训练成本。这凸显了符号推理在增强鲁棒性方面的有效性，而且无需依赖大规模 Transformer 主干网络或增强流程。

5.2 推理时计算比较

前几节主要讨论了鲁棒性和准确性，现在我们将比较我们的神经符号模型与对抗训练以及基于 Transformer 的 LNL-MoEx[18] 在推理时计算成本方面的差异。如前所述，我们的模型采用修改后的 ResNet18 主干网络backbone，后接符号头。该设置能够在扰动样本和干净样本上，同时对类别预测和符号一致性进行推理。每个符号头都是此前所述的轻量级多层感知机。由此产生的推理时计算成本接近于标准 ResNet18 加上额外符号头的成本，远低于基于 Transformer 的防御方法。相比之下，经过对抗训练的模型（例如经过 PGD 训练的 ResNet18）只需进行一次前向传播，且不涉及符号推理，因此成本略低，但不具备可解释性。与此同时，LNL-MoEx 采用基于 ViT 的骨干网络，并结合矩交换逻辑和局部性感知约束，导致其推理时复杂度显著更高。

尽管依赖符号推理，我们的模型仍显著比基于 Transformer 的防御方法更高效。如表 4 所示，我们的方法实现了实用的权衡——在保持强大鲁棒性和逻辑驱动的可解释性的同时，相较于标准 ResNet18，计算量仅小幅增加。由于交通标志识别系统是自动驾驶流水线中不可或缺的组成部分

表 4. 不同架构的推理时计算成本对比。我们的方法以适度的开销兼顾了推理能力与鲁棒性。

方法	骨干网络	准确率（PGD）	推理成本
对抗训练	ResNet18	38.65%	极低
LNL-MoEx[18]	ViT	59.4%	High
NeuroShield	ResNet18 + 多层感知机	56.0%	Low

，其在资源有限且计算预算受限的设备上的部署是一个关键考量。因此，降低推理时计算开销对于实现实际应用和广泛落地至关重要。一个很有前景的方向是探索诸如提前退出机制或模型量化等技术，正如以往针对边缘部署的神经网络优化研究所提出的[28,29]，以进一步降低计算复杂度，同时保持符号鲁棒性和可解释性。

6 结论

本文介绍了**NeuroShield**，这是一种通过整合符号规则监督来增强深度学习模型对抗鲁棒性和可解释性的神经符号框架。通过在 GTSRB 数据集上开展大量实验，我们证明，与标准对抗训练和最先进的基于 Transformer 的防御方法相比，我们的模型表现更优，同时所需训练轮数更少，且无需数据增强。我们的 PGD-神经符号模型在快速梯度符号法和 PGD攻击下实现了可接受的准确率，同时保持了在干净数据集上的准确率。此外，我们的方法在鲁棒性与推理能力之间实现了平衡，且推理时计算开销适中，因此有望应用于现实场景。未来，我们计划进一步调整框架，使其适用于边缘设备和能源受限环境，例如自动驾驶汽车所使用的设备和环境；为此，我们将探索轻量级骨干网络、提前退出机制和量化推理。另一个方向是将符号推理从静态规则检查拓展到链式逻辑推理，使模型能够将领域知识应用于动态驾驶场景，并从基础规则中生成新的符号规则。这样一来，神经符号系统便能在不确定性下辅助更高层次的决策，推动可信人工智能在安全关键应用中的发展。

数据可用性

本研究中的实验使用了 GTSRB 数据集，该数据集可通过以下 Kaggle 仓库公开获取。该数据集提供了实验所用的原始训练图像和测试图像。本研究使用的所有对抗数据（快速梯度符号法和投影梯度下降）均由配套代码仓库中实现的算法即时生成，没有创建或托管任何额外数据集。用于生成这些数据和复现实验的完整代码可在 GitHub 代码仓库中获取。其他实验材料（例如训练后的模型检查点和日志）可向通讯作者提出合理请求后获取。

参考文献

1. Wang, W., Yang, Y. & Wu, F. 面向数据与知识驱动的人工智能：神经符号计算综述。 《*IEEE* 模式分析与机器智能汇刊》(2024年)。**2.** Sze, V., Chen, Y.-H., Yang, T.-J. & Emer, J. S. 深度神经网络的高效处理：教程与综述。 《*IEEE* 会刊》 **105**, 2295–2329 (2017年)。**3.** Badreddine, S., Garcez, A. d., Serafini, L. & Spranger, M. 逻辑张量网络。 《人工智能》 **303**, 103649 (2022年)。**4.** Taha, I. A. & Ghosh, J. 人工神经网络的符号解释。 《*IEEE* 知识与数据工程汇刊》 **11**, 448–463 (1999年)。**5.** Chakraborty, A., Alam, M., Dey, V., Chattopadhyay, A. & Mukhopadhyay, D. 对抗攻击与防御措施综述。《*CAAI* 智能技术汇刊》**6**, 25–45 (2021年)。**6.** Khoshavi, N. 等 Shieldenn：面向容错深度神经网络架构的在线加速框架。载于*2020*年 第*57*届 *ACM/IEEE* 设计自动化会议 *(DAC)*, 1–6 (IEEE, 2020年)。**7.** Dong, H. 等 神经逻辑机。 *arXiv* 预印本 *arXiv:1904.11694* (2019年)。**8.** Shindo, H., Dhami, D. S. & Kersting, K. 神经符号前向推理。 *arXiv* 预印本 *arXiv:2110.09383*(2021年)。

9. Xu, J., Zhang, Z., Friedman, T., Liang, Y. & Broeck, G. 一种结合符号知识的深度学习语义损失函数。载于国际机器学习大会，5502–5511 (PMLR, 2018年)。10. Xie, X., Kersting, K. & Neider, D. 深度神经网络的神经符号验证。arXiv 预印本 arXiv:2203.00938 (2022年)。11. Goodfellow, I. J., Shlens, J. & Szegedy, C. 解释并利用对抗样本。arXiv 预印本 arXiv:1412.6572 (2014年)。12. Madry, A., Makelov, A., Schmidt, L., Tsipras, D. & Vladu, A. 面向能够抵御对抗攻击的深度学习模型。arXiv 预印本 arXiv:1706.06083 (2017年)。13. Athalye, A., Carlini, N. & Wagner, D. 混淆梯度带来虚假的安全感：规避针对对抗样本的防御措施。载于国际机器学习大会，274–283 (PMLR, 2018年)。14. Carlini, N. & Wagner, D. 对抗样本并非易于检测：绕过十种检测方法。载于第十届 ACM 人工智能与安全研讨会论文集，3–14 (2017年)。15. Cohen, J., Rosenfeld, E. & Kolter, Z. 通过随机平滑实现认证对抗鲁棒性。载于国际机器学习大会，1310–1320 (PMLR, 2019年)。16. Ahmed, K. 等 Pylon：一种用于约束学习的 PyTorch 框架。载于 NeurIPS 2021 竞赛与演示赛道，319–324 (PMLR, 2022年)。17. Stallkamp, J., Schlipsing, M., Salmen, J. & Igel, C. 德国交通标志识别基准：一项多类别分类竞赛。载于2011年国际神经网络联合会议，1453–1460 (IEEE, 2011年)。18. Manzari, O. N., Kashiani, H., Dehkordi, H. A. & Shokouhi, S. B. 具有局部性归纳偏置和特征归一化的鲁棒 Transformer。《工程、科学与技术国际期刊》38, 101320 (2023年)。19. Wang, W. 等 金字塔视觉 Transformer：一种无需卷积的通用密集预测骨干网络。载于IEEE/CVF 国际计算机视觉大会论文集，568–578 (2021年)。20. Han, K. 等 Transformer 中的 Transformer。《神经信息处理系统进展》34, 15908–15919 (2021年)。21. Yuan, L. 等 Tokens-to-Token 视觉变换器：在 ImageNet 上从头训练视觉 Transformer。载于IEEE/CVF 国际计算机视觉大会论文集，558–567 (2021年)。22. Mao, X. 等 面向鲁棒视觉 Transformer。载于IEEE/CVF 计算机视觉与模式识别大会论文集，12042–12051 (2022年)。23. Liu, Z. 等 Swin Transformer：采用移位窗口的层次化视觉 Transformer。载于IEEE/CVF 国际计算机视觉大会论文集，10012–10022 (2021年)。24. Hendrycks, D. 等 鲁棒性的多重面貌：对分布外泛化的批判性分析。载于IEEE/CVF 国际计算机视觉大会论文集，8340–8349 (2021年)。25. Yun, S. 等 CutMix：利用可定位特征训练强分类器的正则化策略。载于IEEE/CVF 国际计算机视觉大会论文集，6023–6032 (2019年)。26. Hendrycks, D. 等 AugMix：一种提升鲁棒性与不确定性的简易数据处理方法。arXiv 预印本 arXiv:1912.02781 (2019年)。27. Kim, J.-H., Choo, W. & Song, H. O. Puzzle Mix：利用显著性与局部统计量实现最优 mixup。载于国际机器学习大会，5275–5285 (PMLR, 2020年)。28. Shafiee Sarvestani, A., Tabrizchi, S., Sehatbakhsh, N. & Roohi, A. RL-SEP：基于强化学习的智能退出点选择，以延长能量采集系统的寿命。载于第23届 ACM 嵌入式网络传感器系统会议论文集，638–639 (2025年)。29. Solanki, N., Tabrizchi, S., Sohrabi, S., Schmidt, J. & Roohi, A. ATM-Net：面向能量采集边缘智能的自适应终止与多精度神经网络。arXiv 预印本 arXiv:2502.09822 (2025年)。

资金支持

本研究部分得到了美国国家科学基金会（NSF）资助，资助编号为 2504839、2447566 和 2448133。

作者贡献声明

A. Sh. 提出了主要构想，并负责最初的研究设计。A. Sh. 和 J. S. 实施了仿真、开展了分析，并撰写了初稿。J. S. 进一步完善并解读了研究结果。A. R. 参与了概念的构建，对仿真和分析进行了批判性审阅，讨论并解读了研究发现，修改并校对了稿件，同时负责监督整个研究。

声明

利益冲突

作者声明不存在利益冲突。

其他信息

通信 及材料索取事宜请联系 Arman Roohi。