arXiv:2302.01056v3 [cs.CV] 9 Nov 2023

超越预训练特征：噪声图像建模提供对抗防御

Zunzhi You[1], **Daochang Liu**[1],**Bohyung Han**[2],**Chang Xu**[11] 悉尼大学计算机科学学院[2] 首尔国立大学电气与计算机工程系和 IPAI
zyou5834@uni.sydney.edu.au, daochang.liu@sydney.edu.au,
bhhan@snu.ac.kr, c.xu@sydney.edu.au

摘要

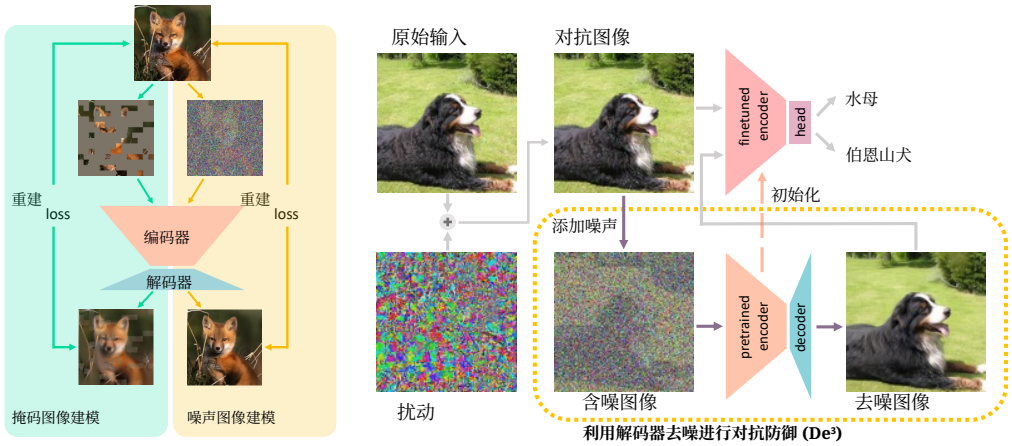
近年来，掩码图像建模（MIM）取得了长足进展，已成为自监督视觉表征学习的主流框架。然而，和大多数深度神经网络方法一样，MIM 预训练模型仍易受对抗攻击影响，这限制了其实际应用，而这一问题鲜少得到研究关注。本文探讨了这一强大的自监督学习范式如何为下游分类器提供对抗鲁棒性。在研究过程中，我们发现，噪声图像建模（NIM）是 MIM 的一种简单变体，它采用去噪作为前置任务；即使图像受到严重破坏，也能出乎意料地很好地重建含噪图像。受此观察启发，我们提出了一种对抗防御方法，称为 De[3], , 通过利用预训练解码器进行去噪。借助 De[3], , NIM 能够在提供预训练特征之外进一步增强对抗鲁棒性。此外，我们还进行了一项简单改进：从随机分布中采样噪声尺度超参数，使该防御方法在准确率与鲁棒性之间实现更优且可调节的权衡。实验结果表明，得益于有效的去噪能力，NIM 在对抗鲁棒性方面优于 MIM。此外，NIM 提供的防御性能与对抗训练相当，同时还具备额外的可调节优势。源代码和模型可在 https://github.com/youzunzhi/NIM-AdvDef 获取。

1 引言

微调经过预训练的深度神经网络，是迄今为止机器学习取得成功的关键。经过大规模数据集充分预训练的深度神经网络，可以为各种下游任务提供良好的初始化，并提升其性能。在这一范式中，自监督学习发挥着至关重要的作用，因为这种方法使网络能够利用海量无标签数据。随着强大且可扩展的 Transformer 架构视觉模型的出现， [14, 21], 掩码图像建模（MIM）近年来迅速发展，并成为视觉特征预训练的新主流范式， [38, 17, 8, 26], 这得益于其在自然语言处理（NLP）领域遵循掩码语言建模 [12, 4] 取得的成功。从概念上讲，掩码图像建模背后的思路很简单。在预训练阶段，系统会随机按图像块对无标签训练图像进行掩码处理，然后将其输入编码器-解码器架构；解码器则尝试根据编码器嵌入的特征恢复原始图像。

尽管 MIM 等预训练方法提升了下游视觉任务的标准性能，但微调后模型的对抗鲁棒性并未增强。面对难以察觉却经过恶意精心构造的扰动时，模型依然容易受到攻击，这将严重限制 MIM 在现实世界中的部署，尤其是在安全关键应用中。尽管现有最先进的方法，

第37届神经信息处理系统大会（NeurIPS 2023）。



(a) (b)

图1：(a) NIM 框架与 MIM 框架的比较。除了退化步骤外，这两个框架几乎完全相同。(b) 所提出的 De^3 对抗防御示意图。De^3 首先向受攻击图像添加一定程度的噪声，然后使用通过 NIM 框架预训练的模型对图像进行去噪。

对抗训练 [22, 23], 能够训练出对抗鲁棒模型，但众所周知，其计算成本高昂 [22]，并且可能导致模型在干净数据上的准确率大幅下降 [34],，从而削弱 MIM 预训练特征的实用性。此外，与标准训练相比，实现对抗鲁棒泛化所需的样本复杂度显著更高，这使得标签效率在实践中尤为重要。鉴于这些问题，我们提出以下问题：如何在实现标准准确率的同时，通过当前最先进的生成式预训练范式提供对抗鲁棒性?

在探讨这个问题时，我们发现这一范式有两个显著特点：(1) 除遮蔽图像块之外，其他退化方式，例如放大、模糊或在频域中进行遮蔽，也可有效用于自监督学习的前置任务 [33, 37]。(2) 对 MIM 的尝试性理解表明，它隐式地学习了遮挡不变特征 [19]。受这些观察启发，我们探索了噪声图像建模（NIM），这是 MIM 的一种变体，它将去噪用作前置任务，并学习编码噪声不变特征（图1 (a)）。
在探索过程中，我们发现 NIM 的去噪图像质量出奇地高，即使输入的是噪声非常强的图像也是如此（图3（e）和（g））。然而，在 MIM 的常见做法中，只有编码器会被保留下来以供进一步微调，而执行图像重建的解码器则会被丢弃。因此，NIM 强大的去噪能力将被浪费；我们认为，这种能力可用于去除对抗扰动。为此，我们提出了一种名为 De^3 的对抗防御方法：**De**noising by **De**coder for Adversarial **De**fense（通过解码器去噪以实现对抗防御）。该方法首先向对抗样本添加一些高斯噪声，然后尝试重建原始图像，如图1（b）所示。借助 De^3, , NIM 不仅为微调模型提供预训练特征，还提供对抗防御。我们进一步提出，在预训练过程中对控制退化程度的超参数进行随机化，使 NIM 提供的防御能够在标准准确率和对抗鲁棒性之间实现可调节的权衡，从而进一步提升其在实践中的适用性。

总而言之，本研究的贡献如下：

- 我们提出了一种名为 De^3 的新方法，利用 NIM 模型强大的去噪能力，使其能够在预训练特征之外提供对抗攻击防御。
- 不同于现有 MIM方法将控制退化程度的超参数设为全局固定值，我们提出从随机分布中对其进行采样，从而使防御实现灵活的准确率-鲁棒性权衡，并可在测试时进行调节。

• 大量实验表明，我们的 NIM预训练能够学习到与 MIM 同样优良的视觉特征，同时还具有优势，因为它能有效提升微调模型抵御强力对抗攻击的鲁棒性。

2 相关工作

2.1 视觉领域的自监督预训练

从无标注图像中学习视觉表征，是机器学习和计算机视觉领域的一个研究方向，近年来受到越来越多的关注。掩码语言建模（MLM）在大规模语言模型预训练中取得巨大成功 [12,4, 20], 而在最近之前，视觉模型预训练的主流框架一直是对比学习（CL） [6, 18,7] 。随着适用于视觉任务的 Transformer 大量涌现 [14, 5, 21],，视觉领域中与 MLM 对应的方法——MIM，已成为新的主流自监督预训练范式 [26, 13, 17, 38]。通过掩码并预测图像块，编码器能够学习提取有意义的视觉特征。

除了将掩码预测用作前置任务外，也有研究考察了采用其他退化方法的效果。Tian 等人 [33] 研究了五种方法，包括放大、缩小、失真、模糊和去色，并发现这些方法的表现均优于监督式预训练。此外，Xie 等人 [37] 探索了频域中的掩码操作，并证明超分辨率和去噪等类似的低级任务也能取得有竞争力的表现。这些研究启发我们探索噪声图像建模框架。最近，Fang 等人 [15] 也对生成式预训练框架中的掩码操作进行了改进：输入图像由一个额外的小型可训练 BEiT 生成器进行退化 [2]。然而，我们并非仅将生成式预训练框架用于获取预训练特征，而是进一步提出利用该框架增强对抗鲁棒性。

2.2 对抗鲁棒性

大约十年前，研究人员发现深度神经网络容易受到难以察觉的对抗扰动的影响 [32, 3], 。此后，如何增强模型抵御此类攻击的鲁棒性便成为一个活跃的研究领域，而更强大的威胁模型也在不断发展 [16, 25, 22, 10]。最成功的方法之一是对抗训练（AT）：在训练过程中引入对抗样本，以增强模型的对抗鲁棒性 [16, 22, 39]。尽管对抗训练有效，但这类方法面临诸多挑战，限制了其适用性。例如，由于训练时必须针对每张图像搜索有效攻击，计算成本可能过高。此外，对抗训练在提高鲁棒性的同时，会牺牲标准准确率 [34, 27]。尽管 Wang 等 [36] 通过模型条件训练方法缓解了这一问题，使鲁棒性与准确率之间的权衡能够进行原位校准，但该方法进一步增加了训练成本，而且仅适用于卷积神经网络。

另一条研究路线旨在防御输入空间中的对抗攻击，而不是像对抗训练（AT）一样在模型空间中进行防御。这些方法通常使用额外的生成器，从遭受攻击的图像中恢复干净图像，例如生成对抗网络（GAN） [29] 或扩散模型（Diffusion Models） [24]。由于其训练不像对抗训练那样依赖威胁模型，这些方法通常能够防御未见过的威胁。然而，它们的性能通常无法与对抗训练方法相媲美 [10]。我们的 De^3 方法与这些方法不同，因为我们的防御能力由预训练提供，无需额外的生成模型。因此，获得防御能力所付出的成本还带来了预训练特征，并提升了微调效果。更重要的是，这项工作的主要目标是展示 NIM 相较于掩码图像建模（MIM）在对抗鲁棒性方面的优势，而不是提出一种新颖的对抗防御方法。

3 噪声图像建模

3.1 初步介绍：掩码图像建模

我们首先简要回顾 MIM 预训练框架。对于从训练数据集 $\mathbf{\Omega}$中采样的每张图像 $\mathbf{x} \in \mathbb{R}^{H\times W\times C}$，其中$H$、 W和 C 分别表示高度、宽度和通道维度，MIM 首先在掩码比例

超参数 γ 的控制下随机采样一个二值掩码 $\mathbf{M}$，使得 $\sum_i^N m_i/N = \gamma$，其中二值掩码 $\mathbf{M}$中的每个元素 $m_i \in 0,1$ 对应图像中的一个像素，而 N 表示像素总数。然后，掩码图像表示为

$$\hat{\mathbf{x}}_{MIM} = \mathbf{x} \odot (\mathbf{1} - \mathbf{M}), \tag{1}$$

其中，$\odot$ 表示逐元素乘积。之后，MIM 使用其编码器 f_θ 和解码器 g_ϕ，分别由 θ 和 ϕ参数化，来预测重建图像。实际上，MIM 的编码器f_θ 通常由视觉变换器（ViTs） [14] 或其变体 [21], 构成，并采用按块掩码。最后，MIM 使用某种距离度量 $\mathcal{D}$（$\cdot$，$\cdot$），例如 ℓ_1 损失，来衡量被掩码图像块的重建质量。总的来说，MIM 的目标函数如下:

$$\min_{\theta,\phi} \mathop{\mathbb{E}}_{\mathbf{x}\in\mathbf{\Omega}} \mathcal{D}(g_\phi(f_\theta(\hat{\mathbf{x}}_{MIM})) \odot \mathbf{M}, \mathbf{x} \odot \mathbf{M}). \tag{2}$$

3.2 噪声图像建模：学习噪声不变特征

图1 (a)展示了 MIM 与 NIM 预训练过程之间的相似性。作为 MIM 的一种变体，NIM 共享许多组成部分，包括网络架构设计、重建目标和训练技术。唯一的区别在于，NIM 通过添加高斯噪声使原始图像退化，如

$$\hat{\mathbf{x}}_{NIM} = \mathbf{x} + \sigma\epsilon, \quad \epsilon \sim \mathcal{N}(0, \mathbf{I}), \tag{3}$$

其中，σ 是控制退化程度（即噪声尺度）的参数。NIM 的目标函数为

$$\min_{\theta,\phi} \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathbf{\Omega}} \mathcal{D}(g_\phi(f_\theta(\hat{\mathbf{x}}_{NIM})), \mathbf{x}). \tag{4}$$

由于 MIM 学习预测被掩码的区域，在 NIM 下预训练的编码器-解码器能够对受损输入图像进行去噪。请注意，我们使用的“噪声”一词与经典去噪自编码器（DAE） [35]中的含义不同。在我们的语境中，它指的是被添加 到图像中的高斯噪声，而 DAE 则通过将随机选定的像素替换为 0 来破坏输入，这实际上等价于掩码。

继从遮挡不变特征学习的角度理解 MIM 的相关工作之后 [19], ，我们介绍 NIM 如何学习噪声不变特征。假设存在一个由 $g'_{\phi'}$ 参数化的解码器网络 ϕ'，它几乎可以将 f_θ 嵌入的特征还原为其输入图像:

$$g'_{\phi'}(f_\theta(\mathbf{x})) \approx \mathbf{x}. \tag{5}$$

因此，式（4）中的 NIM 目标函数可以改写为

$$\min_{\theta,\phi} \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathbf{\Omega}} \mathcal{D}(g_\phi(f_\theta(\hat{\mathbf{x}}_{NIM})), g'_{\phi'}(f_\theta(\mathbf{x}))). \tag{6}$$

通过定义一种新的相似度度量

$$\mathcal{D}'_{\phi,\phi'}(z_1, z_2) \triangleq \mathcal{D}(g_\phi(z_1), g'_{\phi'}(z_2)), \tag{7}$$

式（6）可进一步简化为

$$\min_{\theta,\phi} \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathbf{\Omega}} \mathcal{D}'_{\phi,\phi'}(f_\theta(\hat{\mathbf{x}}_{NIM}), f_\theta(\mathbf{x})). \tag{8}$$

因此，NIM 的目标函数可以理解为最小化原始图像$\mathbf{x}$ 与含噪图像 $\hat{}\mathbf{x}_{NIM}$之间特征嵌入的距离，这意味着由 f_θ 学习到的特征嵌入应当具有噪声不变性。

4 De^3：通过解码器去噪实现对抗防御

4.1 De^3 的形式化表述

尽管掩码图像建模在表征学习方面效果显著，但学习鲁棒的预训练特征仍然颇具挑战。微调模型很容易受到微小对抗扰动的影响，因而难以应用于安全关键任务。意识到这一问题后，我们探索了 NIM 的潜力，并发现 NIM 模型的重建能力出乎意料地强。例如，图3 (e) 和 (g) 展示了由含噪图像重建得到的图像。我们观察到，大多数

语义在重建过程中都得以保留，即使图像噪声过大、人类无法识别也是如此。然而，掩码图像建模预训练的常见做法仅使用编码器来初始化下游模型，而将具有出色重建能力的解码器直接舍弃。与此同时，我们认为 NIM 解码器卓越的去噪能力可用于增强下游模型的对抗鲁棒性，并提出了一种简单的防御技术，称为 De^3：**De**用解码器去噪以实现对抗**防御**。

如图1 (b)所示，De^3 在下游任务的测试时运行：首先向输入图像添加高斯噪声，然后使用预训练解码器对图像进行去噪，具体如下：

$$y = c_\psi(g_\phi(f_\theta((\mathbf{x}' + \sigma_{De^3}\epsilon)))), \tag{9}$$

其中，c_ψ 是给定预训练 θ后由 ψ 参数化的微调后的下游网络，$\mathbf{x}'$ 是可能遭受对抗攻击的输入图像，σ_{De^3} 是用于防御的噪声尺度。这种方法背后的动机很直接：由于对抗扰动被限制在较小的 ϵ 范围内，因此很容易被幅度大得多的高斯噪声淹没，随后在解码过程中连同随机高斯噪声一起去除。此外，测试时随机生成噪声，即使在白盒攻击设置下，威胁模型通常也无法完全获取噪声，这使得找到有效攻击方法更加困难。

虽然 NIM 模型的重建结果逼真且误差极小，但重建图像与原始输入图像之间仍存在明显差距。因此，我们提出使用去噪图像微调下游模型；用于微调的训练样本 c_ψ 也经过先添加噪声再去噪的处理，以模拟测试时遇到的条件。微调的目标函数表述如下：

$$\min_{\psi} \mathop{\mathbb{E}}_{\mathbf{x}\in\mathbf{\Omega}} \mathcal{L}(c_\psi(g_\phi(f_\theta(\mathbf{x} + \sigma_{ft}\epsilon))), \bar{y}), \tag{10}$$

其中，$\mathcal{L}$ 是损失函数，$\bar{y}$ 是x的真实标签，σ_{ft} 是微调过程中使用的噪声尺度。请注意，在微调过程中，θ 和 ϕ 是固定的。

4.2 通过在预训练中采样随机 σ 实现可调节的权衡

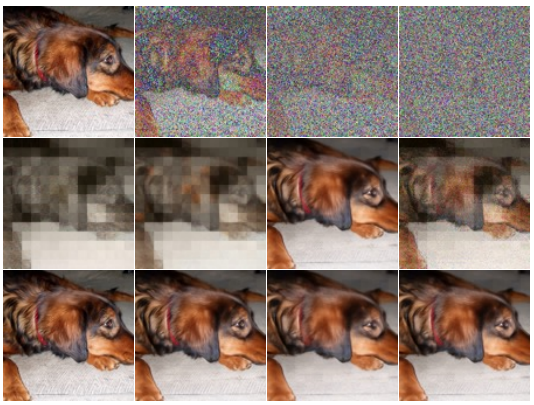
直观而言，通过调整 σ_{De^3}，所提出的 De^3 方法应能在干净准确率与对抗鲁棒性之间实现可调节的权衡。增大 σ_{De^3} 会使对抗扰动更容易被淹没并随后移除，从而增强对抗鲁棒性。与此同时，退化越严重，重建质量就越差，导致干净准确率降低。然而，我们发现这一假设在实践中未必成立。图 2 (a) 展示了在训练和测试期间对输入图像施加不同噪声水平时，重建质量会下降。我们进一步展示了重建损失[1]随噪声水平的变化情况，如图 2 (b) 所示，并在全局设定 $\sigma = 25$、75 和 150 的预训练模型中观察到了相同现象。这些观察表明，如果模型在训练期间仅接触单一噪声水平，就难以将其去噪能力泛化到具有不同噪声水平的图像。

为了解决这一问题，我们提出从随机分布中采样噪声水平参数 σ，而不是像现有 MIM 方法那样将其全局固定。图 2 展示了，使用随机选取的 σ 进行预训练的模型能够高质量地重建输入图像，即使面对不同噪声尺度下的受损图像也不例外，前提是噪声尺度不过大。因此，我们可以在 NIM 模型的准确率与鲁棒性之间找到良好的权衡。

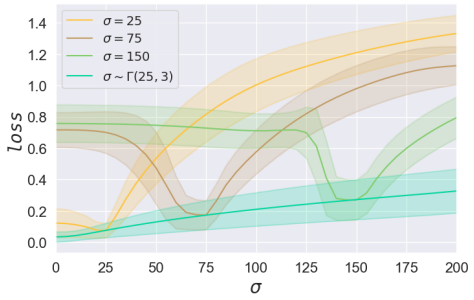
5 实验结果

在本节中，我们通过实验证明，NIM 作为一种自监督预训练方法，能够通过 De^3 方法带来对抗鲁棒性。我们在第 5.1 节介绍实验设置，并在第 5.2 节将 NIM 与掩码图像建模进行比较。随后，在第 5.3 节中，我们将所提出的基于 NIM 的对抗防御 De^3 与对抗训练方法进行比较。最后，第 5.4 节通过消融研究探讨预训练中的 σ 如何影响模型性能。更多实验结果和细节请参见补充材料。

[1]损失越低，重建效果越好。



(a) 比较采用全局和随机 sigma 的 NIM预训练模型对不同噪声水平图像的重建效果。顶部：噪声水平分别为 $\sigma = 0$ （原始图像）、50、75、100 的含噪图像。中部：使用预训练时 σ 全局设为 75 的 NIM 模型重建的图像。底部：使用预训练时 σ 从 Γ(25,3) 随机采样的 NIM 模型重建的图像。



(b) 采用 σ 全局设为 25、75、150，或从 Γ(25,3) 采样进行预训练的模型，其重建损失与噪声水平 σ 之间的关系。图中也显示了标准差。

图 2：定性和定量比较表明，采用全局设定的 σ进行 NIM 预训练时，只能很好地重建噪声水平接近预训练中 σ 的含噪图像；而采用随机 σ 进行 NIM 预训练时，则能在合理范围内对任意 σ 进行良好重建。

5.1 实验设置

数据集与骨干网络我们在 ImageNet-1K [11] 数据集上进行了所有实验。预训练和微调使用训练集（128 万张图像），评估使用验证集（50,000 张图像）。在主要实验中，我们采用 ViT-Base（ViT-B/16） [14] 作为骨干网络。其他骨干网络的结果见补充材料。

威胁模型为评估对抗鲁棒性，我们采用半径为 $\epsilon = 4/255$ 的非定向 ℓ_∞ 有界攻击，这是 ImageNet 对抗鲁棒性研究中最常用的设置。我们考虑三种常见的白盒攻击：单步攻击 FGSM [16], 、多步攻击 PGD [22], ，以及最先进的攻击 AutoAttack（AA） [10]。对于 PGD，我们将步数设为 $n = 10$ ，步长设为 $\alpha = 2/255$。对于 AA，由于我们的防御是随机化防御，因此使用其“rand”版本。为节省计算成本，我们在 RobustBench 从 ImageNet1K 验证集中选取的 5000 张图像子集上应用 AA 攻击 [9]。

训练实现我们采用两种具有代表性的 MIM 方法——MAE [17] 和 Sim-MIM [38], ——作为基准 MIM 方法。我们只修改退化部分，训练 NIM-MAE 和 NIM-SimMIM：将前置任务从掩码预测（使用全局设定的掩码比例）改为去噪（使用随机采样的 σ），其余超参数和训练技术均遵循原论文。默认模型使用 $\sigma \sim \Gamma$ 进行预训练（25,3），并在 $\sigma \sim U$（0, 30）的去噪图像上进行微调。所有模型均预训练 800 个训练轮次，然后微调 100 个训练轮次。

5.2 比较 NIM 与掩码图像建模

在表1中，我们将 NIM 模型与 MIM基线进行比较。首先，我们表明，在没有任何对抗防御的情况下，NIM预训练分类器的准确率略低，但比 MIM预训练分类器更具鲁棒性。因此，我们认为，即使在无法使用防御模型的情况下，NIM 仍是一种简单而有效的自监督视觉学习框架，值得在未来进一步研究。

表1：MIM 和 NIM-MIM 在 ImageNet-1K 上的干净样本准确率与鲁棒性准确率（%）对比，展示了是否采用所提出的 De^3 防御。 †De^3 指将 De^3 适配到 MIM 方法的防御方法（即由解码器进行去掩码以实现对抗防御）。De^{3}后面的数字表示防御中使用的退化程度，其中 γ 为掩码比例， σ 为噪声尺度。采用 De^3 进行防御的 NIM-MAE 模型，使用来自预训练模型的去噪图像进行微调。

模型	掩码自编码器 [17]		NIM-MAE			SimMIM [38]		NIM-SimMIM		
De^3	None	γ=0.75	None	σ=70	σ=140	None	γ=0.6	None	σ=70	σ=140
干净样本	**83.05**	44.96	82.58	78.68	70.69	**83.62**	43.05	82.76	76.23	68.76
FGSM	28.29	36.38	31.66	**53.66**	53.12	30.38	34.29	31.90	**51.84**	51.67
PGD-10	0.25	11.58	0.31	34.61	**39.82**	1.13	7.39	2.02	33.41	**37.45**
AA	0.00	2.58	0.00	23.24	**33.70**	0.00	1.82	0.00	21.70	**32.46**



图3：掩码自编码器和 NIM-MAE 预训练模型的重建定性结果。(a) 原始图像。(b) 75% 掩码图像。(c) 掩码自编码器预训练模型对 (b) 的重建图像。(d) 添加了 σ =70 高斯噪声的图像。(e) 我们的 NIM-MAE 预训练模型对 (d) 的重建图像。(f) 添加了 σ =140 高斯噪声的图像。(g) 我们的 NIM-MAE 预训练模型对 (f) 的重建图像。所示图像从 ImageNet-1K 验证集中随机选取。

借助 NIM 预训练模型提供的 De^3 防御，下游微调分类器对所有白盒攻击都表现出很强的鲁棒性。从数值上看，当高斯噪声的 σ 为 70（或 140）时，添加的高斯噪声幅度约为对抗噪声的 $70/4 = 17.5$ （或 35）倍。因此，对抗噪声可以被高斯噪声淹没，并随高斯噪声一起被 De^{3}去除。由此，采用 De^3 的 NIM 模型在抵御 FGSM 攻击时，其鲁棒性优于未采用防御的 MIM对应模型，最高提升 25.37%，而干净准确率仅下降 4.37%。De^3 防御对 PGD-10 和 AA 等强攻击也有效；这些攻击几乎能以 100% 的成功率使原始模型失效。

相比之下，掩码图像建模并不能从防御中获益。第2列和第7列展示了使用掩码图像建模预训练模型进行对抗防御的结果。这里，我们将 De^3 防御调整为适用于掩码图像建模模型：首先随机掩蔽一些图像块，然后通过预训练编码器和解码器重建掩码图像，其中掩码比例与预训练时相同（掩码自编码器为 75%，SimMIM 为 60%）。结果表明，掩码图像建模预训练模型无法提供有效的对抗防御：干净准确率大幅下降，而鲁棒性仅略有提升。图3展示了分别使用掩码自编码器和 NIM-MAE 重建掩码图像与含噪图像的定性结果。值得注意的是，掩码自编码器的重建效果远差于 NIM-MAE，即使 σ 为 140，图像已无法被人类辨认，情况依然如此。我们还通过计算整个 ImageNet-1K验证集上原始图像与重建图像之间的峰值信噪比（PSNR，数值越高越好），对重建质量进行了定量评估。结果显示，当掩码比例为 75% 时，掩码自编码器的 PSNR 为 20.88；而 NIM-MAE 在 σ =70 时为 27.43，在 σ =140 时为 24.9。

表 2：采用 De^3 防御的对抗训练 ViT-Base 模型与 NIM-MAE 模型在 ImageNet-1K 上的干净样本准确率和鲁棒性准确率（%）比较。

预训练	防御	干净样本	FGSM	PGD-10	AA
从头开始	对抗训练	56.97	39.89	24.25	19.50
MAE	对抗训练	69.67	51.80	39.65	34.56
NIM-MAE	对抗训练	69.28	52.77	**39.97**	**34.84**
NIM-MAE	De^3 ($\sigma = 140$)	**70.69**	**53.12**	39.82	33.70

5.3 比较 NIM+De^3 与对抗训练

为进一步了解采用 De^3 防御的 NIM 模型的有效性，本节将其与对抗训练方法进行比较。对视觉变换器进行对抗训练并非易事，适用于卷积神经网络的技术和超参数可能并不适用于视觉变换器，甚至无法带来益处 [1, 30]。这里，我们采用近期一项研究 [23]提出的视觉变换器对抗训练方案。我们进行 20 轮次的 PGD-5 对抗训练，使训练时间与我们的方法相近，并相应调整学习率调度。表2展示了从头开始进行对抗训练，以及使用掩码自编码器和 NIM-MAE 预训练模型进行对抗训练的结果。结果表明，掩码自编码器和 NIM-MAE 均有助于视觉变换器的对抗训练。与对抗模型相比，我们采用σ=140 De^3 防御的 NIM-MAE 模型具有相当的鲁棒性，且干净准确率略高。

对抗训练模型往往会过度牺牲干净准确率。在我们的实验中，MAE 模型在自然图像上的性能下降了 13.38%。虽然可以通过在目标函数中为标准损失项设置更大的权重来缓解这一问题 [39], ，但调整权衡的代价是必须重新开始另一个训练过程。相比之下，我们的 De^3 防御支持在测试时调节权衡，也就是说，我们无需训练另一个网络，就能在鲁棒性与准确率之间进行权衡（反之亦然）。在测试时增大防御中添加的噪声幅度，可以让预训练去噪模型淹没并去除更多对抗扰动，但代价是重建质量变差、语义信息损失增加（图3（e）与（g）对比），从而提高鲁棒准确率并降低干净准确率（图4）。例如，我们的模型对 PGD-10 攻击的鲁棒性可从接近 0 提高到 31.25%，而干净准确率（79.37%）仅降低 2%。因此，在实际应用中，我们的防御可以根据具体情境的需求，调整其在干净图像和对抗图像上的性能。例如，在自动驾驶中，当车辆进入恶劣环境时，对抗鲁棒性的需求可能会提高；而在更安全的场所，则会优先考虑较高的干净准确率。

在撰写本文期间，出现了一些针对视觉变换器进行对抗训练的较新研究 [28, 31], ，这些研究通过更精细的实现和更长的训练时间，显著提升了视觉变换器在 ImageNet-1K 上的对抗鲁棒性。尽管我们采用 De^{3}方法的 NIM 未能达到这些现有工作的对抗鲁棒性水平，但我们想强调的是，本研究的目的并非与最先进方法竞争。相反，我们旨在向研究界展示，NIM 可以作为一种颇具前景且有优势的自学习范式，用于提升对抗鲁棒性。

5.4 σ 预训练中的 σ

本节将介绍消融研究，以评估预训练中高斯噪声尺度参数 σ 的随机分布如何影响微调模型的性能。我们采用掩码自编码器作为掩码图像建模的默认实现，并在评估鲁棒准确率时随机选取 5,000 张图像，以降低实验开销。

首先，我们展示未采用防御的模型在干净样本上的准确率和鲁棒准确率。图5左侧子图显示，NIM预训练模型在干净图像上能够取得有竞争力的准确率，尤其是在将退化程度超参数设为或集中在适当值时（即，$\sigma = 75, \sigma \sim \Gamma(25, 3)$，$\sigma \sim U(50, 100)$）。另一方面，只要 σ 的方差不是太大（例如，$\sigma \sim U(0, 250)$，$\sigma \sim \Gamma(3, 25)$），大多数 NIM 模型的对抗鲁棒性都高于 MAE。这一现象与第3.2节中关于 NIM 学习噪声不变特征的理解一致。

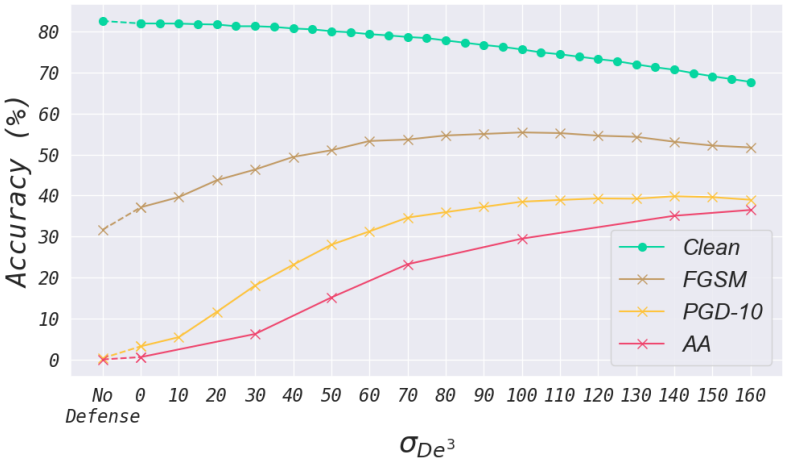


图4：使用不同 σ 的 De^3 NIM-MAE 的干净准确率和鲁棒准确率。随着 σ 增大，干净准确率下降，鲁棒准确率上升，从而可以在测试时调节权衡。

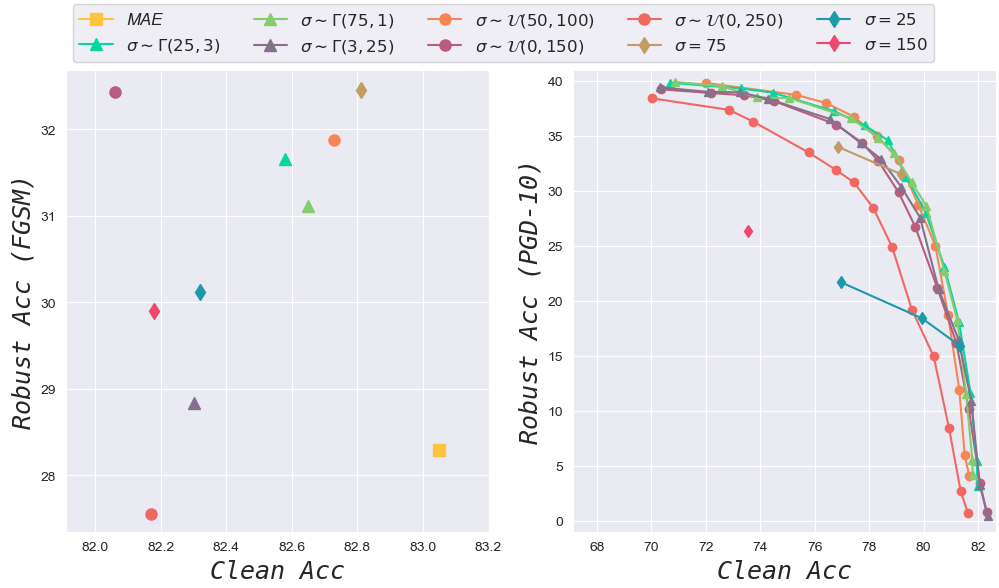


图5：预训练时使用不同 σ 的 MAE 和 NIM-MAE 的性能。左：未采用 De^3 防御的 MAE 和 NIM-MAE 微调模型在干净图像以及 FGSM 攻击的对抗图像上的准确率。右：采用 De^3 防御的 NIM-MAE 模型在干净准确率与抵御 PGD-10 攻击的鲁棒准确率之间的帕累托前沿。通过在防御中使用不同的 σ （从 0 到 160，每次增加 10）来测量帕累托前沿。

在图5右侧子图中，我们展示了使用不同 NIM-MAE 预训练模型时 De^3 的经验帕累托前沿，即图中的点表示通过在防御中使用不同的 σ，模型在干净样本准确率和鲁棒准确率之间所能达到的最佳权衡。我们首先观察到，使用全局设定的 σ 时，模型的帕累托前沿收敛到很少几个点，因为使用预训练时所用值以外的 σ 会同时降低干净样本准确率和鲁棒准确率。同时，结果表明，从伽马分布中采样的 σ 通常比从均匀分布中采样的 σ 实现更好的准确率-鲁棒性权衡。在所有模型中，$\sigma \sim \Gamma(25, 3)$表现最佳，因为其取值集中在适中范围且方差合适。

6 结论

本文研究了噪声图像建模（NIM），这是一种简单的掩码图像建模变体，其前置任务从掩码预测改为去噪。我们发现，NIM 即使面对噪声强烈的图像，也展现出卓越的图像重建能力，因此尝试利用这种去噪能力去除对抗扰动。我们提出了一种简单的方法，使 NIM 模型不仅能提供预训练特征，还能提供对抗防御。为了在测试时进一步实现可调节的权衡，我们从随机分布中采样噪声级别超参数，而不是像 MIM 那样对其进行全局固定。大量实验结果表明，就对抗鲁棒性而言，NIM 凭借强大的去噪能力优于 MIM。我们还表明，NIM 提供的对抗防御性能可与近期的一种对抗训练方法相媲美，并且由于其准确率-鲁棒性权衡可调，具有更高的灵活性和适用性。我们希望本研究能促使学界探索 NIM 及 MIM 的其他变体，从而充分释放生成式视觉预训练的潜力。

致谢

本研究部分由澳大利亚研究委员会项目 DP240101848 和 FT230100549 资助，另有部分经费来自韩国政府（科学技术信息通信部）资助的国家研究基金会和信息与通信技术规划评估院项目（编号：2022R1A2C3012210、2021-0-01343）。

参考文献

[1] Yutong Bai、Jieru Mei、Alan L Yuille 和 Cihang Xie。Transformer 比卷积神经网络更鲁棒吗? 神经信息处理系统进展，34：26831–26843，2021年。[2] Hangbo Bao、Li Dong 和 Furu Wei。BEiT：用于图像Transformer的BERT预训练。*arXiv*预印本 *arXiv:2106.08254*，2021年。[3] Battista Biggio、Igino Corona、Davide Maiorca、Blaine Nelson、Nedim Šrndi´c、Pavel Laskov、Giorgio Giacinto 和 Fabio Roli。测试时针对机器学习的规避攻击。载于数据库中的机器学习与知识发现：欧洲会议，*ECML PKDD 2013*，布拉格，捷克共和国，*2013*年*9*月*23*日至*27*日，论文集，第三部分 *13*，第387–402页。施普林格，2013年。[4] Tom Brown、Benjamin Mann、Nick Ryder、Melanie Subbiah、Jared D Kaplan、Prafulla Dhariwal、Arvind Neelakantan、Pranav Shyam、Girish Sastry、Amanda Askell 等。语言模型是少样本学习器。神经信息处理系统进展，33：1877–1901，2020年。[5] Hanting Chen、Yunhe Wang、Tianyu Guo、Chang Xu、Yiping Deng、Zhenhua Liu、Siwei Ma、Chunjing Xu、Chao Xu 和 Wen Gao。预训练图像处理Transformer。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第12299–12310页，2021年。[6] Ting Chen、Simon Kornblith、Mohammad Norouzi 和 Geoffrey Hinton。用于视觉表征对比学习的简单框架。载于国际机器学习会议，第1597–1607页。PMLR，2020年。[7] Xinlei Chen、Saining Xie 和 Kaiming He。自监督视觉Transformer训练的实证研究。载于*IEEE/CVF*国际计算机视觉会议论文集，第9640–9649页，2021年。[8] Yabo Chen、Yuchen Liu、Dongsheng Jiang、Xiaopeng Zhang、Wenrui Dai、Hongkai Xiong 和 Qi Tian。SDAE：自蒸馏掩码自编码器。载于欧洲计算机视觉会议，第108–124页。施普林格，2022年。[9] Francesco Croce、Maksym Andriushchenko、Vikash Sehwag、Edoardo Debenedetti、Nicolas Flammarion、Mung Chiang、Prateek Mittal 和 Matthias Hein。RobustBench：标准化的对抗鲁棒性基准。*arXiv*预印本 *arXiv:2010.09670*，2020年。[10] Francesco Croce 和 Matthias Hein。通过多种无参数攻击集成，可靠地评估对抗鲁棒性。载于国际机器学习会议，第2206–2216页。PMLR，2020年。[11] Jia Deng、Wei Dong、Richard Socher、Li-Jia Li、Kai Li 和 Li Fei-Fei。ImageNet：大型分层图像数据库。载于*2009* 年 *IEEE*计算机视觉与模式识别会议，第248–255页。IEEE，2009年。[12] Jacob Devlin、Ming-Wei Chang、Kenton Lee 和 Kristina Toutanova。BERT：用于语言理解的深度双向Transformer预训练。*arXiv*预印本 *arXiv:1810.04805*，2018年。[13] Xiaoyi Dong、Jianmin Bao、Ting Zhang、Dongdong Chen、Weiming Zhang、Lu Yuan、Dong Chen、Fang Wen 和 Nenghai Yu。PECO：用于视觉Transformer的BERT预训练的感知码本。*arXiv*预印本 *arXiv:2111.12710*，2021年。

[14] Alexey Dosovitskiy、Lucas Beyer、Alexander Kolesnikov、Dirk Weissenborn、Xiaohua Zhai、Thomas Unterthiner、Mostafa Dehghani、Matthias Minderer、Georg Heigold、Sylvain Gelly 等。一张图像胜过16×16个词：用于大规模图像识别的Transformer。arXiv预印本 arXiv:2010.11929，2020年。[15] Yuxin Fang、Li Dong、Hangbo Bao、Xinggang Wang 和 Furu Wei。用于自监督视觉预训练的损坏图像建模。载于第十一届国际学习表征会议，2023。[16] Ian J Goodfellow、Jonathon Shlens 和 Christian Szegedy。解释并利用对抗样本。arXiv预印本 arXiv:1412.6572，2014年。[17] Kaiming He、Xinlei Chen、Saining Xie、Yanghao Li、Piotr Dollár 和 Ross Girshick。掩码自编码器是可扩展的视觉学习器。载于IEEE/CVF计算机视觉与模式识别会议论文集，第16000–16009页，2022年。[18] Kaiming He、Haoqi Fan、Yuxin Wu、Saining Xie 和 Ross Girshick。用于无监督视觉表征学习的动量对比。载于IEEE/CVF计算机视觉与模式识别会议论文集，第9729–9738页，2020年。[19] Xiangwen Kong 和 Xiangyu Zhang。通过学习遮挡不变特征理解掩码图像建模。arXiv预印本 arXiv:2208.04164，2022年。[20] Yinhan Liu、Myle Ott、Naman Goyal、Jingfei Du、Mandar Joshi、Danqi Chen、Omer Levy、Mike Lewis、Luke Zettlemoyer 和 Veselin Stoyanov。RoBERTa：一种鲁棒优化的BERT预训练方法。arXiv预印本 arXiv:1907.11692，2019年。[21] Ze Liu、Yutong Lin、Yue Cao、Han Hu、Yixuan Wei、Zheng Zhang、Stephen Lin 和 Baining Guo。Swin Transformer：采用移位窗口的层次化视觉Transformer。载于IEEE/CVF国际计算机视觉会议论文集，第10012–10022页，2021年。[22] Aleksander Madry、Aleksandar Makelov、Ludwig Schmidt、Dimitris Tsipras 和 Adrian Vladu。迈向能够抵御对抗攻击的深度学习模型。载于国际学习表征会议，2018年。[23] Yichuan Mo、Dongxian Wu、Yifei Wang、Yiwen Guo 和 Yisen Wang。当对抗训练遇上视觉Transformer：从训练到架构的实践方法。载于 NeurIPS，2022年。[24] Weili Nie、Brandon Guo、Yujia Huang、Chao wei Xiao、Arash Vahdat 和 Animashree Anandkumar。用于对抗净化的扩散模型。载于国际机器学习会议，第16805–16827页。PMLR，2022年。[25] Nicolas Papernot、Patrick McDaniel、Somesh Jha、Matt Fredrikson、Z Berkay Celik 和 Ananthram Swami。对抗环境下深度学习的局限性。载于2016年IEEE欧洲安全与隐私研讨会（EuroS&P），第372–387页。IEEE，2016年。[26] Zhiliang Peng、Li Dong、Hangbo Bao、Qixiang Ye 和 Furu Wei。BEiT v2：采用向量量化视觉分词器的掩码图像建模。arXiv预印本 arXiv:2208.06366，2022年。[27] Aditi Raghunathan、Sang Michael Xie、Fanny Yang、John Duchi 和 Percy Liang。理解并缓解鲁棒性与准确率之间的权衡。载于国际机器学习会议，第7909–7919页。PMLR，2020年。[28] Sylvestre-Alvise Rebuffi、Francesco Croce 和 Sven Gowal。重新审视对抗训练中的适配器。载于第十一届国际学习表征会议，2023。[29] Pouya Samangouei、Maya Kabkab 和 Rama Chellappa。Defense-GAN：利用生成模型保护分类器免受对抗攻击。arXiv预印本 arXiv:1805.06605，2018年。[30] Rulin Shao、Zhouxing Shi、Jinfeng Yi、Pin-Yu Chen 和 Cho-Jui Hsieh。视觉Transformer的对抗鲁棒性。arXiv预印本 arXiv:2103.15670，2021年。[31] Naman D Singh、Francesco Croce 和 Matthias Hein。重新审视ImageNet上的对抗训练：跨威胁模型的架构、训练与泛化。arXiv预印本 arXiv:2303.01870，2023。[32] Christian Szegedy、Wojciech Zaremba、Il ya Sutskever、Joan Bruna、Dumitru Erhan、Ian Goodfellow 和 Rob Fergus。神经网络的奇妙特性。arXiv预印本 arXiv:1312.6199，2013年。[33] Yunjie Tian、Lingxi Xie、Jiemin Fang、Mengnan Shi、Junran Peng、Xiaopeng Zhang、Jianbin Jiao、Qi Tian 和 Qixiang Ye。超越掩码：揭开视觉Transformer基于令牌的预训练之谜。arXiv预印本 arXiv:2203.14313，2022年。[34] Dimitris Tsipras、Shibani Santurkar、Logan Engstrom、Alexander Turner 和 Aleksander Madry。鲁棒性可能与准确率相悖。arXiv预印本 arXiv:1805.12152，2018年。[35] Pascal Vincent、Hugo Larochelle、Yoshua Bengio 和 Pierre-Antoine Manzagol。利用去噪自编码器提取并组合鲁棒特征。载于第25届国际机器学习会议论文集，第1096–1103页，2008年。[36] Haotao Wang、Tianlong Chen、Shupeng Gui、TingKuei Hu、Ji Liu 和 Zhangyang Wang。一次性对抗训练：免费实现鲁棒性与准确率的原位权衡。神经信息处理系统进展，33：7449–7461，2020年。[37] Jiahao Xie、Wei Li、Xiaohang Zhan、Ziwei Liu、Yew Soon Ong 和 Chen Change Loy。用于自监督视觉预训练的掩码频率建模。arXiv预印本 arXiv:2206.07706，2022年。[38] Zhenda Xie、Zheng Zhang、Yue Cao、Yutong Lin、Jianmin Bao、Zhuliang Yao、Qi Dai 和 Han Hu。SimMIM：掩码图像建模的简单框架。载于IEEE/CVF计算机视觉与模式识别会议论文集，第9653–9663页，2022年。[39] Hongyang Zhang、Yaodong Yu、Jiantao Jiao、Eric Xing、Laurent El Ghaoui 和 Michael Jordan。理论上严谨的鲁棒性与准确率权衡。载于国际机器学习会议，第7472–7482页。PMLR，2019年。