基于推理阶段的对抗视觉调优方法

张家明 桑基韬 于 剑
（北京交通大学计算机与信息技术学院 北京 100091）

摘 要 近年来，大规模预训练视觉-语言模型在图像描述、视觉问答和图像检索等任务中展现出卓越的性能。然而，这些模型在安全性方面存在显著的脆弱性，尤其容易受到几乎不可见的对抗噪声的攻击。对抗噪声通过在输入图像中加入人眼几乎不可察觉的扰动，使得模型发生错误。这种脆弱性在实际应用中带来了深度学习模型的安全性挑战，特别是在处理敏感信息的任务中。尽管对抗训练已被证明可以有效提升模型的对抗鲁棒性，但由于其计算复杂度较高，难以直接应用于大规模的视觉-语言模型。为应对这一挑战，本文提出了一种基于推理阶段的对抗视觉调优方法（Adversarial Inference-time Visual Prompt Tuning，AI-VPT），首次在推理阶段针对视觉模态进行提示调优，旨在增强视觉编码器的对抗鲁棒性。AI-VPT通过学习视觉嵌入向量，在推理过程中与对抗图像嵌入对齐，优化视觉表示以削弱对抗性噪声的影响。具体而言，AI-VPT在对抗样本上生成多种增强视图，通过信息熵筛选低熵视图以保留有效信息，从不同角度减弱对抗性干扰，从而进一步提高模型的对抗防御能力。相比于传统的对抗训练技术Adversarial Training，AI-VPT减少了92.9%的时间成本，显著降低了计算开销，尤其适用于大规模预训练视觉-语言模型。经过在六个高分辨率视觉数据集上的广泛测试，AI-VPT展现出了显著的优势，在ViT-B/16和ViT-L/14架构上相对于现有的对抗提示调优方法分别提升了26.1%和18.5%的对抗鲁棒性。

关键词 深度学习；视觉-语言模型；对抗防御；提示学习；图像识别
中图法分类号 TP311 **DOI号** 10.11897/SP.J.1016.2025.01443

Adversarial Inference-Time Visual Prompt Tuning

ZHANG Jia-Ming SANG Ji-Tao YU Jian
（School of Computer Science and Information Technology, Beijing Jiaotong University, Beijing 100091）

Abstract In recent years, large-scale pre-trained vision-language models have demonstrated exceptional performance in tasks such as image captioning, visual question answering, and image retrieval. However, these models exhibit significant vulnerabilities in terms of security, particularly being susceptible to attacks involving almost imperceptible adversarial noise. Adversarial noise induces errors in the model by adding perturbations to the input image that are barely detectable to the human eye. This vulnerability presents a security challenge for deep learning models in practical applications, especially in tasks involving sensitive information. Although adversarial training has been proven to effectively improve the adversarial robustness of models, its high computational complexity makes it difficult to directly apply to large-scale vision-language models. To address this challenge, we propose an adversarial inference-time visual prompt tuning method (AI-VPT), which introduces the concept of prompt tuning for the visual modality during the inference stage to enhance the adversarial robustness of the visual encoder. AI-VPT learns visual embedding vectors and aligns them with adversarial image embeddings during inference, optimizing visual representations to mitigate the effects of adversarial noise. Specifically, AI-VPT generates multiple enhanced views on adversarial samples and selects low-entropy views through information entropy filtering to preserve meaningful information, weakening

收稿日期：2024-12-02；在线发布日期：2025-04-09。**张家明**，博士，主要研究领域为多媒体分析、计算机视觉和多模态模型。E-mail：jiamingzhang@bjtu.edu.cn。**桑基韬**（通信作者），博士，教授，中国计算机学会（CCF）会员，主要研究领域为多媒体分析、数据挖掘和深度学习。E-mail：jtsang@bjtu.edu.cn。**于 剑**，博士，教授，中国计算机学会（CCF）会士，主要研究领域为机器学习、数据挖掘和深度学习。

adversarial interference from various perspectives. As a result, the method significantly improves the model's defense against adversarial attacks. Compared to traditional adversarial training techniques, AI-VPT reduces training time by 92.9%, greatly decreasing computational overhead, and is especially suitable for large-scale pre-trained vision-language models. Extensive testing on six high-resolution visual datasets demonstrates the substantial advantages of AI-VPT, achieving a 26.1% and 18.5% improvement in adversarial robustness on ViT-B/16 and ViT-L/14 architectures, respectively, compared to AdvPT(Adversarial Prompt Tuning, AdvPT).

Keywords deep learning; vision-language models; adversarial defense; prompt learning; image recognition

1 引 言

近年来,大规模预训练视觉-语言模型(Vision-Language Models, VLMs)成为深度学习领域的研究焦点[1-3]。这些模型通过利用海量数据来建立视觉与语言模态之间的联系,从而展现出强大的泛化能力。随着多模态学习技术的不断发展,越来越多的大规模预训练模型被开源,这进一步促进了其在诸如图像描述、视觉问答、图像检索等任务中的应用[4-6]。

尽管大规模视觉-语言模型在各种应用场景中表现优异,但其对安全性问题依然存在显著脆弱性。研究表明,与传统的视觉模型类似,大规模视觉-语言模型也易受到精心设计且几乎不可见的对抗性噪声(Adversarial Noise)攻击[7]。对抗性噪声指的是通过对输入图像施加微小扰动,使得模型在处理这些图像时产生错误,而这些扰动对人类视觉系统来说几乎无法察觉[8-10]。这种脆弱性为深度神经网络在实际应用中的安全性带来了严峻挑战,尤其是在涉及敏感信息的任务中。

在提升深度神经网络对抗鲁棒性的方法中,对抗训练[11]被广泛认为是提升深度神经网络对抗攻击鲁棒性的有效方法。对抗训练通过在每次训练迭代中生成对抗样本并用于更新视觉编码器,以此增强模型的对抗鲁棒性。但是,由于对抗训练过程计算复杂度高,导致其难以直接应用于大规模视觉-语言模型,尤其是那些拥有数百万甚至数十亿参数的模型。为了解决上述问题,Zhang等人提出了一种对抗提示调优(Adversarial Prompt Tuning)的技术[12]。该技术通过学习向量来建模文本提示,并将干净文本嵌入与对抗图像嵌入对齐,从而提高模型的对抗鲁棒性。尽管这种方法降低了传统对抗训练的成本,但它仅聚焦于文本编码器而忽视了视觉编码器的作用,这使得对抗提示调优技术的潜力未能得到充分发挥。

为了解决这一缺陷,我们因此提出了一种新的方法,称为基于推理阶段的对抗视觉调优(Adversarial Inference-time Visual Prompt Tuning, AI-VPT),以在视觉编码器层面进一步增强大规模视觉-语言模型的对抗鲁棒性。AI-VPT 弥补了现有对抗提示调优方法的不足,首次将提示调优的思想引入到视觉模态。具体而言,AI-VPT 通过将视觉嵌入作为可学习向量嵌入到视觉编码器中,并在推理时通过与对抗图像嵌入对齐来优化这些向量,从而提高模型对抗攻击的鲁棒性。给定一个对抗图像,AI-VPT 首先通过多种数据增强技术生成多个视图,这些增强视图从不同角度对对抗噪声进行了降解。接着,通过确保模型在各个增强视图上输出的预测一致性,AI-VPT 能够找到一个最优的方向来削弱对抗噪声,并进一步优化视觉嵌入的表征能力。这种方式有效避免了对抗训练过程中计算资源的大幅消耗,因为所有调优过程仅在推理阶段进行,不增加训练阶段的计算开销。相比于传统的对抗训练技术,AI-VPT 减少了92.9%的时间成本,显著降低了计算开销,尤其适用于大规模预训练视觉-语言模型。经过在六个高分辨率视觉数据集上的广泛测试,AI-VPT 展现出了显著的优势,在 ViT-B/16 和 ViT-L/14 架构上相对于基线方法分别提升了26.1%和18.5%的对抗鲁棒性。

本研究的主要贡献包括:

(1)提出了一种名为AI-VPT的新方法,该方法通过在推理阶段进行视觉提示调优,增强了大规模预训练视觉-语言模型,例如CLIP在图像识别任务上的对抗鲁棒性。与传统对抗训练不同,AI-VPT不增加训练阶段的计算负担,使其更适用于大规模模型。

(2)在多个数据集和模型上进行了广泛的实验验证，结果表明本方法的有效性和优越性。

2 相关工作与背景

为了更好地理解本文提出的方法及其重要性，本节首先概述了对抗攻击与防御的基本概念，接着回顾了视觉-语言模型在对抗性环境下的研究进展，以及提示调优技术在对抗防御中的应用情况。最后，我们讨论了现有研究在多模态对抗性防御方面的局限性，并介绍了本文的主要贡献。

2.1 对抗攻击与防御

对抗攻击(Adversarial Attacks)是指通过向输入数据中添加精心设计的小幅度扰动，使得机器学习模型输出错误结果，而这些扰动对于人类观察者来说几乎是不可见的，目前对抗攻击已成为机器学习安全性研究的重要议题[7]。这类攻击主要分为两大类：有目标攻击(Targeted Attacks)和无目标攻击(Non-targeted Attacks)[13-14]。有目标攻击是指攻击者试图使模型产生特定的错误分类，而非目标攻击则只需模型产生错误分类即可，本研究所探讨的对抗攻击属于后者的范围。根据攻击者对模型内部结构和参数的了解程度，对抗攻击可以分为两大类：白盒攻击(White-box Attacks)和黑盒攻击(Black-box Attacks)[15-16]。白盒攻击假设攻击者完全了解目标模型的内部结构和参数值。在这种情况下，攻击者可以利用这些信息来寻找最优的对抗噪声，以最大化攻击效果。常见的白盒攻击方法包括Fast Gradient Sign Method(FGSM)[9]、Projected Gradient Descent(PGD)[11]等。这些方法通常通过梯度信息来指导对抗样本的生成，从而有效地误导模型。相比之下，黑盒攻击假设攻击者对目标模型一无所知，在实际应用场景中，黑盒攻击更为常见，因为它更接近真实世界中的攻击场景[17-18]。黑盒攻击通常通过模拟白盒攻击的过程，采用随机搜索、迁移攻击或其他启发式方法来生成对抗样本。这些技术试图在没有内部模型信息的情况下，通过大量尝试来找到有效的对抗噪声[19-21]。

为了应对对抗攻击带来的威胁，研究人员提出了多种防御策略。其中，对抗训练是最具代表性的防御方法之一。Madry等人提出了一种基于Min-Max优化框架的对抗训练方案，通过在训练过程中引入对抗样本，使得模型能够在面对对抗攻击时保持较高的分类准确率[11]。其后，有更多的基于Min-Max优化框架的防御方法被提出，诸如TRADES[22]、MART[23]等。尽管这些方法在提高模型鲁棒性方面取得了显著成效，但由于其计算复杂度高，对于大规模预训练模型的应用受到了限制。

2.2 大规模预训练视觉-语言模型

大规模预训练视觉-语言模型在广泛的任务中取得了突破性进展，展现出卓越的跨模态学习与推理能力。这些模型能够通过同时处理图像和文本数据，实现高度复杂的语义对齐，成为诸多应用领域的核心技术。通常情况下，大规模预训练视觉-语言模型可以分为两大类：第一类模型基于大型自然语言处理架构，通过在其基础上集成视觉模态来扩展其应用能力。这类模型以语言为主导，视觉模态作为补充，用于增强其多模态推理能力。第二类模型则以CLIP[24]和ALIGN[2]为代表，强调图像和语言模态的平等重要性。这类模型通过大规模自监督学习，构建出联合的图像-语言表征，从而在跨模态任务中展现出更高的泛化能力和鲁棒性[25]。这些模型依托于从海量多模态数据中挖掘到的深层语义关系，能够在没有明确标注的情况下学习到图像和语言之间的紧密关联，使其在多模态检索、图像描述生成等任务中表现出色。

本研究的重点是第二类模型。由于图像编码器在处理对抗样本时表现出一定的脆弱性，因此我们旨在提升模型的对抗鲁棒性，进而提升整体系统的安全性。这一研究为大规模预训练视觉-语言模型在安全敏感场景中的应用提供了支持。

2.3 提示学习

提示学习的概念最早起源于自然语言处理领域，其核心思想是通过调整提示(prompt)来实现模型的微调，而无需直接修改模型参数。与传统的微调方法相比，提示学习通过冻结模型本身，仅对输入的提示进行优化，从而降低微调的复杂度[26]。最初的研究侧重于手工设计提示，但后续工作逐步转向自动学习更加有效的提示，以提升模型在特定任务上的表现。随着该技术的不断发展，提示学习已经被扩展到视觉模型和视觉-语言模型中，旨在通过优化提示提高模型的表现和准确性。例如，CoOp框架首次将提示学习引入视觉-语言领域，通过自动化的提示优化提升了模型的泛化能力[27]。该方法通过学习上下文信息，使模型能够更好地适应不同的视觉-语言任务，减少了对人工设计提示的依赖。TPT则是将提示学习部署在了模型的推理阶段以提升模型的泛化能力，这种方式进一步减少了计算

成本，完全不需要在训练阶段对模型做修改[28]。本文所遵循的技术路线属于后者，但是学习目标有所区别，本文专注于提升模型的对抗鲁棒性。

在对抗鲁棒性方面，Zhang等人率先将提示学习技术应用于对抗防御中，提出了对抗文本提示调优的概念[12]。他们的工作通过学习表示向量来建模文本提示，将干净文本的嵌入与对抗样本的图像嵌入进行对齐，从而有效提升模型的对抗鲁棒性。与传统的对抗训练相比，虽然只利用了文本编码器，但是该方法显著降低了计算开销，同时保持了较高的防御效果。这种创新不仅为提升对抗鲁棒性提供了一条新的路径，也为后续研究提供了重要的启示，表明通过提示调优进行对抗防御是一个具备良好发展前景的研究方向。

通过对现有工作的综述，可以发现，尽管大规模预训练视觉-语言模型在多个任务中表现出色，但在对抗鲁棒性方面仍存在明显的不足。提示学习作为一种新兴技术，不仅提升了模型性能，还为解决对抗鲁棒性问题提供了新的思路。基于此，本研究结合提示学习与对抗训练的理念，提出了一种新方法——基于推理阶段的对抗视觉调优（Adversarial Inference-time Visual Prompt Tuning，AI-VPT），在表1中我们展示了新方法与现有策略相比的优越性。本方法旨在克服现有技术的局限，在不增加训练阶段计算开销的前提下，通过关注视觉编码器，提升大规模预训练视觉-语言模型的对抗鲁棒性。

表1 新方法与现有策略的对比

方法	是否无需重训练模型	是否利用图像编码器	计算速度	对抗鲁棒性提升性能
对抗训练	×	√	慢	高
对抗文本调优	√	×	快	低
对抗图像调优	√	√	**快**	**高**

3 推理阶段的对抗视觉调优

在本章中，我们会详细介绍我们所提出的算法——基于推理阶段的对抗视觉调优（Adversarial Inference-time Visual Prompt Tuning，AI-VPT）。首先，我们会介绍需要用到的先验知识，然后再解释了生成多视图的必要性以及多视图的生成流程，最后，我们介绍了如何进行对抗视觉微调。图1给出了整个方法的流程框架图。

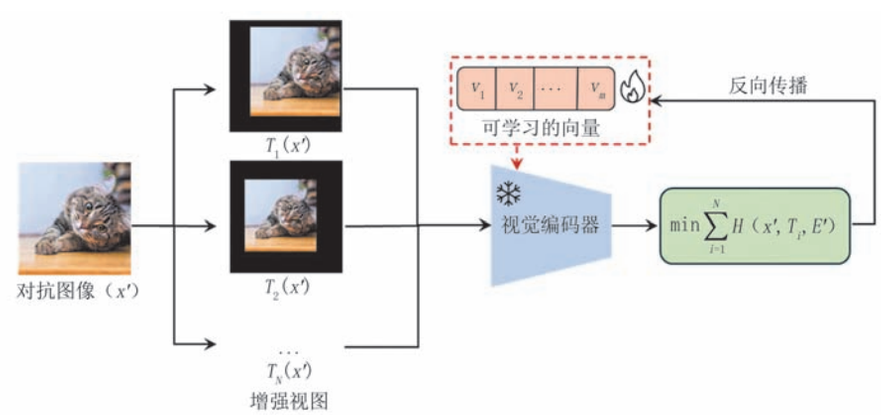

图1 AI-VPT的流程框架图

3.1 先验知识

本文提出的方法基于大规模预训练的视觉-语言模型，重点聚焦于CLIP模型，但该方法具备扩展至更广泛的对比学习框架下的视觉-语言模型的潜力。CLIP模型由图像编码器E和文本编码器T两部分组成。图像编码器负责将输入的视觉信息转化为图像嵌入，以提取不同层次的图像特征。相应地，文本编码器基于Transformer架构，用于将输入的文本转化为文本嵌入，从而实现视觉和语言模态之间的高效对齐。

在训练阶段，CLIP通过使用对比损失来学习统一的嵌入空间，实现视觉模态与语言模态的对齐。通过这种训练方式，模型能够理解图像和文本之间的语义关系，从而在多模态任务中展现出强大的泛化能力。在推理阶段，CLIP可以用于零样本图像识别，即无需针对特定任务进行额外微调，直接通过图像-文本匹配来实现分类。具体的，CLIP通过计算图像与各类别文本描述的相似性来实现分类，比如

输入“a photo of a ＜class＞”的提示语句，其中“＜class＞”为数据集中的某个类别。

设定输入图像为x，其对应的图像嵌入通过图像编码器E生成，记为$e=E(x)$。同时，同时设定文本输入为g，这些提示通过文本编码器Q生成的文本嵌入分别为$w=Q(g)$。模型计算图像与c个类别文本输入的相似性，从而得出预测类别y的概率分布，具体表达式如下：

$$p(y|e)=\frac{\exp(\mathrm{sim}(e,w_y)/\tau)}{\sum_{i=1}^{c}\exp(\mathrm{sim}(e,w_i)/\tau)} \tag{1}$$

其中，sim表示余弦相似度函数，τ为温度参数，用以控制相似度分布的平滑程度。通过这种结构，CLIP模型能够在没有明确标注的情况下，借助对比学习，充分利用大规模的图像和文本数据，实现视觉与语言模态的统一表征。这为模型在各种零样本和跨模态任务中的应用提供了坚实的基础。

接下来引入针对图像编码器的对抗攻击方法。当原始输入图像为x，其中δ表示对抗噪声。经过对抗扰动后的对抗样本为$x'=x+\delta$，当该对抗样本通过图像编码器E处理后，将生成对抗嵌入e'。攻击者通常以破坏图像与文本描述之间的匹配精度为目标，即使对抗嵌入e'与原始图像的嵌入e显著偏离。而为了保证对图像造成的扰动足够的小，在本研究中我们使用了ℓ_∞-范数约束的扰动，其中每个对抗噪声δ满足$\|\delta\|_\infty\leqslant\epsilon$，其中$\epsilon$表示允许的最大扰动幅度。

3.2　多视图生成

推理阶段的数据增强技术可以在一定程度上削弱对抗图像中的对抗性。然而，固定的数据增强方法往往难以达到理想的效果。因此，在本研究中，我们首先对待识别的对抗样本进行多种随机的数据增强，以生成多个不同的视图。这种随机性保证了这些增强视图能够从不同角度削弱对抗性干扰的效果。尽管这些增强方法可以削弱对抗噪声的有效性，但它们也可能同时去除图像中的部分有效信息，导致模型无法对图像进行准确识别。

为了解决这一问题，我们引入了信息熵的概念来度量图像中所包含的有效信息量，其计算公式如下：

$$H(x',T;E)=-\sum_{i=1}^{c} p\Big(\big(y_i\big|T(x')\big)\log p\big(y_i\big|T(x')\big)\Big) \tag{2}$$

其中，T表示所使用的数据增强，c表示类别总数，$p(y_i|T(x'))$表示$T(x')$预测为y_i的概率。信息熵衡量了模型对图像分类的确定性程度，若图像中缺少有效信息，模型的预测值分布会趋于均匀，导致熵值增大。比如，给定一个图像，模型会输出对于该图像的分类结果预测值，当图像中越缺少有效信息，该预测值在每个类上的分布会越平均，熵也就越大。为此，我们首先将对抗图像x'进行K次数据增强，得到与x'对应的多视图$\{x_1',x_2',\cdots,x_K'\}$。接下来，我们会对$K$个视图进行筛选，保留熵最低的前$K_t$个视图用于后续的处理。

3.3　对抗视觉调优

在本研究中，我们利用了 Vision Transformer（ViT）的特性[29]，通过将视觉嵌入v作为可学习的提示嵌入加入到视觉编码器中，从而提升模型对对抗样本x'的识别能力。ViT 具有良好的图像特征提取能力，尤其是在处理高维数据时能够捕捉丰富的语义信息。然而，面对对抗攻击时，模型在对抗样本上往往表现出脆弱性。因此，我们引入可学习的视觉嵌入v，使模型在推理阶段具备更强的鲁棒性。

给定一个输入的对抗图像$x'\in\mathbb{R}^{h\times w\times c}$，ViT会将图像划分为$(h\times w)/p^2$个不重叠的图像补丁，每个补丁的大小为$p\times p$。这些补丁在ViT中作为独立的输入单元，经过一系列线性变换和编码器处理，逐步提取出图像的嵌入表示，每个补丁的维度为$p^2\times c$。在此基础上，我们为每个图像补丁初始化一个视觉嵌入v作为可学习的向量，以提升对对抗图像的处理能力。视觉嵌入v被表示为

$$V=\left\{v_1,v_2,\cdots,v_{(h\times w)/p^2}\right\},\ v_i\in\mathbb{R}^{p^2\times c} \tag{3}$$

针对每个对抗图像x'，我们在推理阶段都会重新进行一次视觉嵌入v的初始化。为了进一步增强模型的鲁棒性，我们对每个对抗图像x'进行多次随机数据增强，从而生成多个视图。如前文所描述的，我们对K个增强视图中的信息熵进行筛选，保留熵最低的前K_t个视图$\{x_1',x_2',\cdots,x_{K_t}'\}$，以用于后续的视觉嵌入优化。通过选择低熵视图，我们能够确保对抗样本经过数据增强后，仍然具有足够的有效信息以供模型识别。

在筛选出的低熵视图基础上，我们通过优化视觉嵌入v来进一步提升模型的对抗鲁棒性。我们定义了一个优化目标，旨在通过减少这些视图的整体熵值来增强模型的识别能力，具体的优化公式如下：

$$\min_{v}\sum_{i=1}^{K_t}H(x',T_i;E') \tag{4}$$

其中，$H(x',T_i)$表示经过数据增强T_i后，视图$x_i'=$

$T_i(x')$的信息熵；E'表示在原始视觉编码器的基础上添加了视觉嵌入v的视觉编码器。通过该优化过程，模型能够在推理阶段动态调整视觉嵌入，从而更好地应对对抗样本，提升对抗攻击下的识别准确性。

4 实验分析

在本章中，我们通过多个实验验证了所提出的AI-VPT方法在对抗鲁棒性、计算效率和可视化效果等方面的表现。该部分全面展示AI-VPT的有效性。我们选择了多个高分辨率视觉数据集，并基于CLIP模型的不同版本进行实验，比较了AI-VPT与基线方法在对抗鲁棒性中表现。同时，为了展示AI-VPT的计算效率优势，我们进行了时间和准确率对比实验。在可视化分析中，我们通过t-SNE技术直观展示了AI-VPT在处理对抗样本时的特征分布情况。最后，针对训练轮数这一超参数，我们分析了其对结果的影响，进一步探讨了最佳参数配置对模型性能的优化作用。

4.1 实验设置

(1)数据集

与之前的工作类似，我们挑选了6个高分辨率的视觉数据集，以及1个图文数据集以进行实验分析，如表2所示。而对于ImageNet的测试集，我们参考了以往关于对抗攻击的相关研究[12]，我们从中随机采样了1000张图像(每个类别采样一张)进行评估。

表2 所使用的数据集以及它们的测试集所包含的图像数量

	Flowers[30]	Pets[31]	Food[32]	SSUN397[33]	UCF101[34]	ImageNet[35]	Flickr30k[36]
数量	2463	3669	30 300	19 850	3783	1000	1000

(2)模型

我们的实验主要基于CLIP模型进行，具体选择了公开可用的ViT-B/16和参数量较大的ViT-L/14版本。在文本输入方面，我们延续了原始CLIP模型的设置，使用了人工设计的提示词(hand-crafted prompts)。例如，对于Flowers数据集，输入的提示语为："a photo of a ＜class＞, a type of flower"，其中＜class＞代表具体的花卉类别。

(3)实现细节

对于视觉嵌入v，我们默认使用AdamW优化器以0.0001的学习率优化128轮。最大扰动幅度ϵ被设置为*16/255*，这也是许多研究中公认的设置。

(4)多视图生成技术

对于所使用的数据增强技术T，我们部署了5种重缩放技术，如表3所示。例如，当尺寸范围为[300，310]时，这张图像会被随机调整到310到331像素之间的某个尺寸，再经过压缩达到310像素。除此之外，我们还部署了一种可微分的JPEG压缩技术。由于对抗噪声通常是高频的图像噪声，而JPEG图像压缩技术是一种保留图像中对人类视觉起决定作用的中低频成分，而去除人类视觉不敏感的高频成分的一种图像处理技术，所以它是多视图技术的一种理想选择。但是由于原生的JPEG图像压缩技术本身是不可为微分的，无法在GPU中进行操作，降低了运行效率，所以我们使用了近似的可微计算代替了不可微分的部分。具体来说，我们实现了一个基于JPEG压缩的可微增强方法。在该方法中，我们首先对输入图像进行2D离散余弦变换(DCT)，然后通过量化操作对其进行压缩，最后通过反量化和逆DCT恢复图像。为确保此过程可微，我们替换了传统JPEG压缩中不可微的部分，采用了可微分的操作，使得图像压缩过程能够在深度学习框架下进行优化和训练。此外，压缩过程中使用的质量因子是随机选择的，这进一步增强了数据增强的多样性和随机性。我们部署了3种质量因子的区间范围，如表4所示。即，我们在8个增强视图种挑选了其中3个熵最低的视图，即$K=8$，$K_r=3$。

表3 重缩放技术所使用的图像尺寸范围

	T_1	T_2	T_3	T_4	T_5
尺寸范围	[300,310]	[310,330]	[310,330]	[310,340]	[330,340]

表4 可微分JPEG压缩技术所使用的质量因子范围

	T_6	T_7	T_8
质量因子范围	[85,90]	[90,95]	[95,100]

4.2 与基线方法的白盒鲁棒性对比

4.2.1 与基于提示调优基线方法的比较结果

在本部分中，我们详细对比了AI-VPT方法与基于提示调优基线方法AdvPT[12](ECCV 2024)在多个高分辨率数据集上的性能表现。实验基于ViT-B/16和ViT-L/14两种模型架构展开，分别在Flowers、Pets、Food101、SUN397、

UCF101和ImageNet数据集上进行实验。我们使用Clean（干净图像下的准确率）、No Defense（PGD-40对抗攻击下的准确率，鲁棒准确率）作为参照，将AdvPT和AI-VPT进行性能比较。表5和表6分别展示了在ViT-B/16和ViT-L/14网络上性能对比结果。

表5　在ViT-B/16网络上AI-VPT与基线方法的白盒鲁棒性（PGD）比较

	Flowers	Pets	Food	SUN397	UCF101	ImageNet
Clean	71.4	89.1	86.1	62.6	66.7	66.1
No Defense	6.4	24.4	14.0	14.7	9.1	6.6
AdvPT	37.4	41.9	38.8	35.7	27.2	19.9
	31.0↑	*17.5*↑	*24.8*↑	*21.0*↑	*18.1*↑	*13.3*↑
Adversarial Training	28.9	9.8	32.4	28.9	5.4	N/A
	22.5↑	*14.6*↓	*18.4*↑	*14.2*↑	*3.7*↓	*N/A*
Fast AT	30.1	12.0	22.9	21.6	13.7	18.4
	23.7↑	*12.4*↓	*8.9*↑	*6.9*↑	*4.6*↑	*11.8*↑
AI-VPT (Ours)	**59.9**	**77.2**	**59.1**	**51.0**	**52.5**	**58.1**
	53.5↑	***52.8***↑	***45.1***↑	***36.3***↑	***43.4***↑	***51.5***↑

表6　在ViT-L/14网络上AI-VPT与基线方法的白盒鲁棒性（PGD）比较

	Flowers	Pets	Food	SUN397	UCF101	ImageNet
Clean	79.3	93.6	91.0	67.6	74.2	72.8
No Defense	20.1	50.3	34.3	27.9	33.9	28.5
AdvPT	56.0	68.7	54.0	44.0	47.9	42.9
	35.9↑	*18.4*↑	*19.7*↑	*16.1*↑	*14.0*↑	*14.4*↑
Adversarial Training	47.7	15.2	40.1	50.7	8.5	N/A
	27.6↑	*35.1*↓	*5.8*↑	*22.8*↑	*25.4*↓	N/A
Fast AT	13.8	10.2	12.7	39.9	10.7	14.1
	6.3↓	*40.1*↓	*21.6*↓	*12.0*↑	*23.2*↓	*14.4*↓
AI-VPT (Ours)	**68.9**	**87.3**	**75.2**	**59.6**	**64.0**	**69.6**
	48.8↑	***37.0***↑	***40.9***↑	***31.7***↑	***30.1***↑	***41.1***↑

如表5所示，在ViT-B/16模型上，AI-VPT方法在大部分数据集上均取得了显著的性能提升（↑代表相对于No Defense的提升）。与AdvPT方法相比，AI-VPT在对抗攻击下的性能提升尤为明显。例如，在Pets数据集上，AI-VPT的准确率为77.2%，相比于AdvPT提升了35.3个百分点。值得注意的是，在所有数据集上，AI-VPT均远远优于没有对抗防御技术（No Defense）的结果。例如，在Flowers数据集上，No Defense的准确率仅为6.4%，而AI-VPT能够提升至59.9%，展现了其强大的对抗鲁棒性。

如表6所示，AI-VPT在ViT-L/14网络上同样表现出了优异的性能（↑代表相对于No Defense的提升）。在Flowers数据集上，AI-VPT的准确率为68.9%，相比于AdvPT提升了12.9个百分点。在Pets数据集上，AI-VPT的准确率为87.3%，几乎接近干净图像下的基线性能（93.6%），显示出其在对抗攻击下依然能够保持较高的识别能力。特别是在ImageNet数据集上，AI-VPT达到了69.6%的准确率，明显优于AdvPT方法的42.9%，充分说明了AI-VPT在大规模图像数据集上的优势。

通过上述实验结果可以看出，AI-VPT方法在多个数据集和两种模型架构下均表现出较高的对抗鲁棒性，显著优于现有的基线方法。特别是在对抗攻击强度较大的情况下，AI-VPT能够有效保持较高的分类准确率。这表明，AI-VPT在处理对抗样本时不仅具有强大的防御能力，还能在多种视觉任务中展现出较好的泛化性能。

4.2.2　与基于对抗训练基线方法的比较结果

为进一步验证方法的普适性，我们在表5和表6中新增了与经典对抗训练方法Adversarial Training[11]和Fast AT[37]的对比。实验表明，传统对抗训练方法在部分数据集上存在显著局限性例如在ViT-B/16的Pets数据集上，Adversarial Training的

鲁棒准确率仅为9.8%，甚至低于未防御（No Defense）的24.4%（下降14.6个百分点），表明其对抗扰动泛化能力不足。类似地，Fast AT在ViT-L/14的Food数据集上仅获得12.7%的准确率，较No Defense下降21.6个百分点。造成这一现象的原因主要有两点：(1)对抗训练的范式在ViT的架构上相比于传统的CNN架构难以收敛，这导致了在某些规模不大的数据集上表现相当差；(2)因为对抗训练范式缺少了原生CLIP模型中的先验知识，其基础表征能力较弱，进而出现了甚至要低于未防御（No Defense）的现象。

相比之下，AI-VPT展现出更稳定的跨数据集适应性。以ViT-L/14的Pets数据集为例，AI-VPT取得87.3%的准确率，较Adversarial Training提升72.1个百分点。特别值得注意的是，在ImageNet等大规模数据集上，Adversarial Training因计算资源限制无法完成训练（N/A），而AI-VPT仍能实现69.6%的鲁棒准确率，验证了其在大规模场景下的可行性。这些结果表明，传统对抗训练方法在面对复杂对抗样本时存在固有缺陷，而AI-VPT通过将对抗学习与视觉提示调优相结合，实现了更优的鲁棒性-泛化性平衡。

4.3 与基线方法的黑盒鲁棒性对比

4.3.1 与基于提示调优基线方法的比较结果

在本部分中，我们评估了AI-VPT方法与基于提示调优基线方法AdvPT的黑盒鲁棒性。与评估白盒鲁棒性的实验一致，基于ViT-B/16和ViT-L/14两种模型架构和Flowers、Pets、Food101、SUN397、UCF101和ImageNet数据集八种数据集上进行实验。为了得到黑盒鲁棒性的量化数据，我们使用了L2T[38]作为黑盒攻击算法，并使用Clean（干净图像下的准确率）、No Defense（L2T对抗攻击下的准确率，鲁棒准确率）作为参照，将AdvPT和AI-VPT进行性能比较。表7和表8分别展示了在ViT-B/16和ViT-L/14网络上性能对比结果。

表7 在ViT-B/16网络上AI-VPT与基线方法的黑盒鲁棒性（L2T）比较

	Flowers	Pets	Food	SUN397	UCF101	ImageNet
Clean	71.4	89.1	86.1	62.6	66.7	66.1
No Defense	52.7	71.7	50.1	38.9	50.6	21.5
AdvPT	62.8	77.9	58.0	55.4	56.6	34.1
	10.1↑	*6.2*↑	*7.9*↑	*16.5*↑	*6.0*↑	*12.6*↑
Adversarial Training	31.1	10.2	52.7	39.9	7.8	N/A
	21.6↓	*61.5*↓	*2.6*↑	*1.0*↑	*42.8*↓	N/A
Fast AT	34.1	17.9	39.4	26.8	9.9	44.7
	18.6↓	*53.8*↓	*10.7*↓	*12.1*↓	*40.7*↓	*23.2*↑
AI-VPT (Ours)	**67.8**	**82.2**	**65.6**	**59.8**	**65.0**	**60.3**
	15.1↑	***10.5***↑	***15.5***↑	***20.9***↑	***14.4***↑	***38.8***↑

表8 在ViT-L/14网络上AI-VPT与基线方法的黑盒鲁棒性（L2T）比较

	Flowers	Pets	Food	SUN397	UCF101	ImageNet
Clean	79.3	93.6	91.0	67.6	74.2	72.8
No Defense	68.5	84.3	57.7	43.2	54.4	35.4
AdvPT	72.3	**90.4**	69.8	58.7	60.0	44.9
	3.8↑	***6.1***↑	*12.1*↑	*15.5*↑	*5.6*↑	*9.5*↑
Adversarial Training	60.7	39.6	66.3	58.3	17.2	N/A
	7.8↓	*44.7*↓	*8.6*↑	*15.1*↑	*37.2*↓	N/A
Fast AT	17.9	14.5	15.0	44.5	13.6	53.6
	50.6↓	*69.8*↓	*42.7*↓	*1.3*↑	*40.8*↓	*18.2*↑
AI-VPT (Ours)	**75.5**	**89.7**	**81.3**	**65.9**	**70.1**	**70.7**
	7.0↑	***5.4***↑	***23.6***↑	***22.7***↑	***15.7***↑	***35.3***↑

从表7可以看出，AI-VPT在多个数据集上的黑盒鲁棒性表现优异。例如，在ViT-B/16的Flowers数据集上，AI-VPT的鲁棒准确率为67.8%，比AdvPT的62.8%高出5.0个百分点。在

Pets数据集上，AI-VPT的鲁棒准确率为82.2%，相比于AdvPT提升了4.3个百分点。特别是在ImageNet数据集上，AI-VPT取得了60.3%的鲁棒准确率，显著高于AdvPT的34.1%。这些结果表明，AI-VPT不仅在白盒攻击下展现出较强的鲁棒性，在黑盒攻击下也能够有效抵御对抗样本的影响。

同样地，表8中的结果也证明了AI-VPT在ViT-L/14网络架构下的黑盒鲁棒性。以Flowers数据集为例，AI-VPT的鲁棒准确率为75.5%，较AdvPT的72.3%提升了3.2个百分点。在较为复杂的ImageNet数据集上，AI-VPT达到了70.7%的鲁棒准确率，比AdvPT高出25.8个百分点。虽然在Pets数据集上略逊于AdvPT 0.7个百分点，但是整体来看，AI-VPT在所有测试数据集和模型架构中均展现出较为稳健的黑盒鲁棒性，证明了其在面对未知对抗样本时的强大防御能力。

4.3.2　与基于对抗训练基线方法的比较结果

为了进一步验证AI-VPT在黑盒环境中的优势，我们将其与基于对抗训练的两种方法——Adversarial Training和Fast AT进行了对比。在L2T黑盒攻击下的实验结果表明，AI-VPT在多个数据集上均超越了传统的对抗训练方法。

如表7所示，AI-VPT在ViT-B/16架构下，尤其在Pets数据集上表现突出，鲁棒准确率为82.2%，相比于Adversarial Training的10.2%提高了72.0个百分点。值得注意的是，与白盒鲁棒性的结果相似，Adversarial Training在该数据集上的鲁棒准确率低于未防御(No Defense)的71.7%，下降了61.5个百分点，显示出其对抗扰动泛化能力的严重不足。类似地，Fast AT在Pets数据集上的表现也未能超过No Defense，鲁棒准确率为17.9%，比No Defense的71.7%下降了53.8个百分点。

在ImageNet数据集上，AI-VPT达到了60.3%的鲁棒准确率，明显高于Fast AT的44.7%。这一差异进一步验证了AI-VPT在大规模数据集下的稳定性和有效性。而传统的对抗训练方法，由于计算资源限制或训练不充分，往往未能达到AI-VPT的鲁棒性水平。

在ViT-L/14架构下，AI-VPT同样表现出了优越的鲁棒性。特别是在UCF101和Food数据集上，AI-VPT的鲁棒准确率分别为70.1%和81.3%，大幅领先于Adversarial Training(分别为17.2%和66.3%)和Fast AT(分别为13.6%和15.0%)。特别是在Food数据集上，Fast AT的鲁棒准确率仅为15.0%，较No Defense的57.7%下降了42.7个百分点，显示了对抗训练范式在ViT架构下的某些小规模数据集上的适应困难。

这些结果表明，AI-VPT相比传统的对抗训练方法(包括Fast AT)在黑盒环境下展现了显著的性能优势。AI-VPT能够在多个数据集和不同的模型架构上维持较高的鲁棒准确率，克服了传统对抗训练方法在训练收敛性和泛化能力上的不足。通过将对抗学习与视觉提示调优相结合，AI-VPT实现了更加稳定的对抗防御，证明了其在复杂黑盒攻击场景中的有效性。

4.4　在多模态任务上的表现

为了证明AI-VPT在多模态任务和超高分辨率数据上的表现，我们选取了Flickr30k数据集，在图像-文本检索(Image-Text Retrieval，IR)任务和图像-文本匹配(Text-Image Retrieval，TR)任务上进行了评估。如表9和表10所示，AI-VPT在这两个任务上都展现出了显著的性能提升。需要注意的是，由于不再是分类任务所以AdvPT无法作为基线方法；而对抗训练会使图像编码器和文本编码器的空间不再对齐，无论是Adversarial Training还是Fast AT，在图像-文本匹配上的表现都几乎为零。

表9　在ViT-B/16网络上的多模态任务(图像-文本检索)的表现

	TR@1	TR@5	TR@10	IR@1	IR@5	IR@10
Clean	81.2	96.4	98.5	62.2	85.6	91.7
No Defense	45.3	77.1	89.9	30.2	68.3	80.5
AI-VPT	**67.5**	**83.2**	**93.3**	**54.9**	**79.4**	**88.6**
(Ours)	***22.2*** ↑	***6.1*** ↑	***3.4*** ↑	***24.7*** ↑	***11.1*** ↑	***8.1*** ↑

表10　在ViT-L/14网络上的多模态任务(图像-文本检索)的表现

	TR@1	TR@5	TR@10	IR@1	IR@5	IR@10
Clean	85.3	97.2	99.2	64.5	87.1	92.1
No Defense	49.6	77.0	91.2	36.5	70.4	82.8
AI-VPT	**70.6**	**85.9**	**95.3**	**59.7**	**82.9**	**90.0**
(Ours)	***21.0*** ↑	***8.9*** ↑	***4.1*** ↑	***23.2*** ↑	***12.5*** ↑	***7.2*** ↑

在ViT-B/16模型上(表9)，AI-VPT在图像-文本检索任务中相较于未防御(No Defense)方法表现出显著优势。例如，在TR@1指标上，AI-VPT的准确率为67.5%，相比No Defense的45.3%提高了22.2个百分点。在IR@1指标上，AI-VPT的准确率为54.9%，相比No Defense的30.2%提升了24.7个百分点。可以看出，AI-VPT不仅在传统的

图像识别任务中展现出优异的鲁棒性，而且能够在多模态任务中进一步提升跨模态匹配的准确性，展示了其良好的多模态泛化能力。

在ViT-L/14模型上（表10），AI-VPT同样展示了更强的多模态任务能力。在TR@1指标上，AI-VPT的准确率为70.6%，较No Defense的49.6%提升了21.0个百分点。在IR@1指标上，AI-VPT的表现同样出色，准确率为59.7%，较No Defense的36.5%提高了23.2个百分点。这些结果表明，AI-VPT能够有效处理跨模态数据，并显著提高图像-文本检索和匹配任务的鲁棒性，证明了其在处理复杂多模态数据集时的优越性。

4.5 计算效率对比分析

4.5.1 与基于提示调优基线方法的比较结果

在本节中，我们首先对比了基于提示调优基线方法AdvPT与我们提出的方法在计算效率上的表现。为了更直观地展示不同方法的训练时间和鲁棒准确率之间的差异，我们在Flowers数据集上使用ViT-B/16网络对比了计算效率的表现。如表11所示，AI-VPT在计算效率方面较现有提示调优方法展现显著优势。尽管AdvPT已通过参数冻结策略将训练时间压缩至1802秒，但AI-VPT进一步将训练耗时降低1.8%（1769秒），同时将鲁棒准确率从56.0%提升至68.9%，实现12.9个百分点的性能飞跃。这种效率提升源于两方面创新：1）将可训练的参数设置在了更具潜力的视觉编码器而不是文本编码器中；2）在测试阶段而非训练阶段进行提示微调，显著提升了效率。实验表明，AI-VPT在保持提示调优轻量化特性的同时，突破了现有方法在效率-鲁棒性权衡上的瓶颈。

表11 AI-VPT与基线方法的效率比较

方法	计算时间/s	鲁棒准确率/%
AdvPT	1802	56.0
Adversarial Training	24 687	47.7
Fast AT	1880	13.8
AI-VPT (Ours)	**1769**	**68.9**

4.5.2 与基于对抗训练基线方法的比较结果

在本节中，我们对比了两种对抗防御方法Fast AT[37]和Adversarial Training[11]与我们提出的方法在计算效率上的表现。结果如表11所示，首先，Adversarial Training方法需要较长的训练时间才能达到相对较高的鲁棒性。经过大约24 687秒的训练，其鲁棒准确率最终达到了47.7%。然而，观察训练过程可见，在前期，特别是在前10 000秒内，鲁棒性提升较为缓慢。相比之下，Fast AT方法的训练时间虽然较短，仅为1880秒，但其鲁棒准确率仅为13.8%，显示出较低的对抗鲁棒性。虽然Fast AT能够有效缩短训练时间，但其在鲁棒性方面的表现明显低于Adversarial Training。

我们提出的方法在计算效率上表现尤为出色。在1769秒内，我们的方法就能够实现68.9%的鲁棒准确率，不仅远超其他两种基线方法，而且展现了更高的计算效率。相比之下，Adversarial Training需要更长时间才能达到类似的鲁棒性，而Fast AT则表现出鲁棒性不足的缺点。

4.6 可视化结果分析

为了更直观地展示我们提出的AI-VPT方法在对抗样本上的有效性，我们采用t-SNE技术将高维数据降至二维，并对比了原生的CLIP模型（No Defense）和AI-VPT两种方法的可视化结果。如图2和图3所示，同类别的数据点使用了相同的颜色，相近类别的数据点使用了相似的颜色，以便观察聚类效果和类别间的区分情况。

通过图2的可视化结果（No Defense）可以发现，未经任何防御的模型在面对对抗攻击时，数据点的分布较为散乱，同一类别的数据点并未形成良好的聚集，甚至出现了不同类别数据点相互混杂的情况。这种分布表明，未加防御的模型，即原生CLIP在对抗攻击下难以保持良好的表征能力，难以对不同类别的样本进行有效的区分。这证明了对抗噪声对模型造成了实质性伤害，显著影响了模型对图像的分类能力。

相比之下如图3所示，我们的方法AI-VPT在处理对抗样本时展现出更优越的聚类效果。正如可视化结果所示，AI-VPT能够有效地将同类数据点聚集在一起，使同类别的数据点形成紧密的簇，同时将不同类别的数据点隔离开。尤其是相近类别的数据点在空间中靠得更近，而类别差异较大的数据点间隔得较远，表明AI-VPT不仅能够增强对抗鲁棒性，还能够保留类别间的语义相似性。这种聚集效果充分说明了AI-VPT在提升模型对抗鲁棒性方面的有效性，使得模型在应对对抗攻击时，依然能够保持较高的类内一致性和类间分离性。这一特性在实际应用中具有重要价值，有助于增强模型在复杂场景下的泛化能力和鲁棒性。

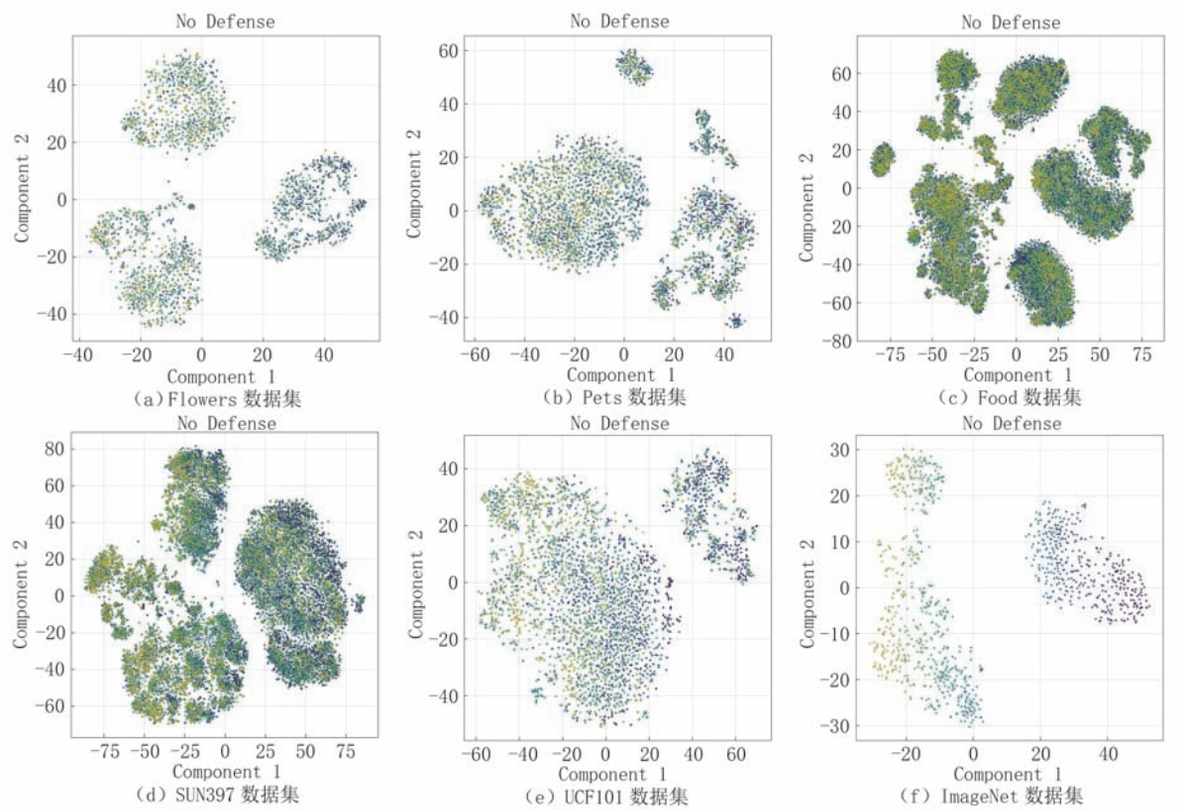

图2　原生CLIP的t-SNE的可视化结果

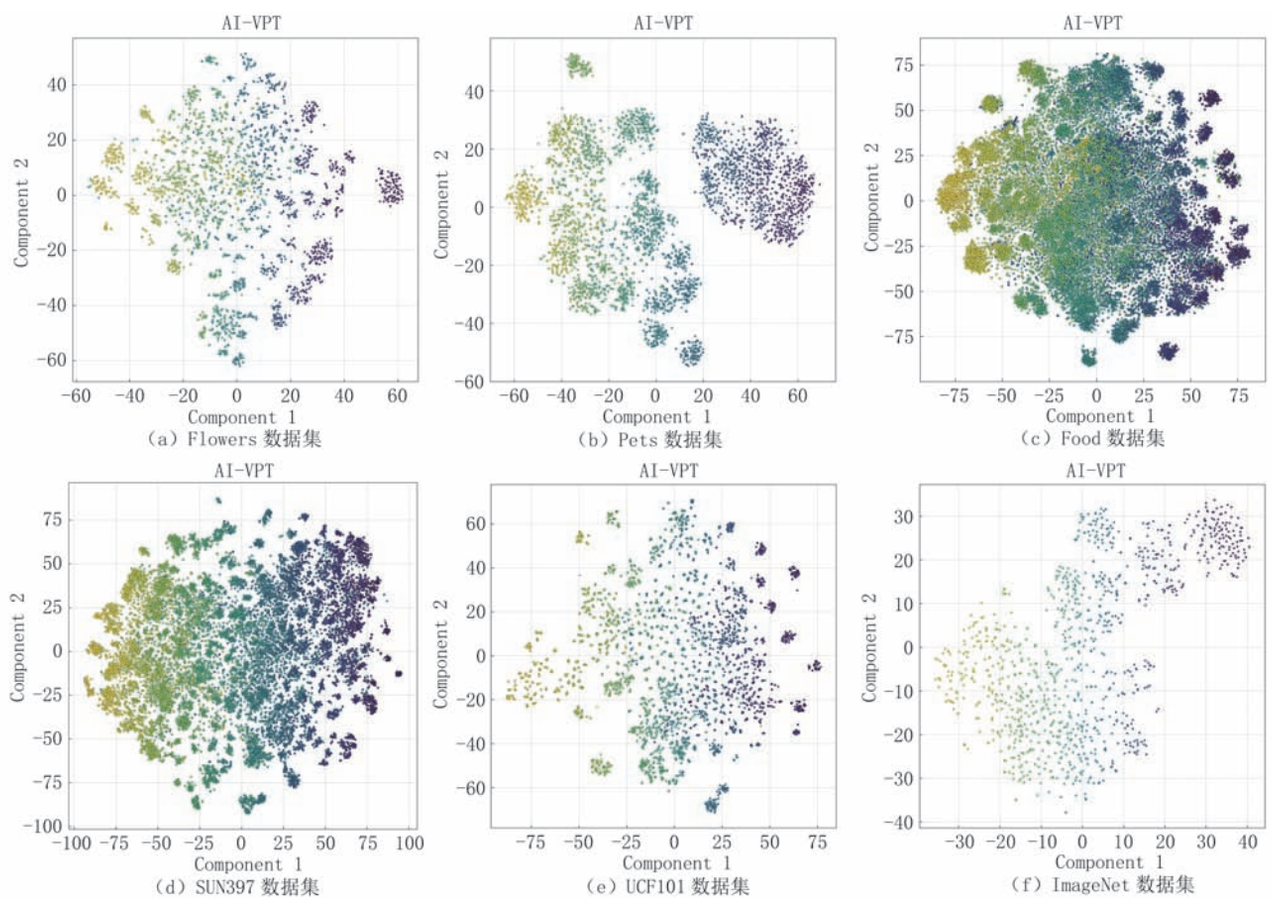

图3　AI-VPT的t-SNE的可视化结果

4.7 参数分析

4.7.1 训练轮数的影响

在本节中，我们分析了训练轮数（epoch）这一超参数对我们方法的影响。为了全面评估AI-VPT在不同轮数下的表现，我们将训练轮数设置在区间[16，128]，步长为4，并记录了各轮数对应的准确率，结果如图4所示。

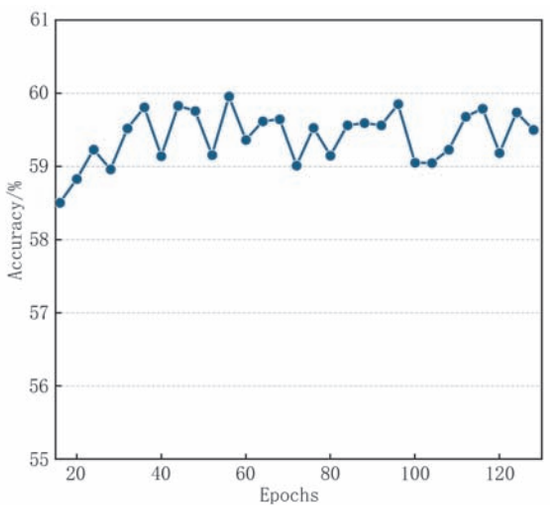

图4 不同训练轮数（Epoch）对AI-VPT的准确率（Accuracy）的影响

从图中可以观察到，模型准确率在训练早期阶段逐步提升，在第36轮时就已经接近最佳值60%，之后的准确率变化较小，呈现出一定的波动但总体维持在59%到60%之间。这表明模型在第36轮时已经基本达到了收敛，进一步增加训练轮数带来的性能提升较为有限。该结果表明，训练轮数控制在36轮左右就足以达到较好的准确率表现，同时节省计算资源和时间。在实际应用中，可以根据任务需求和资源条件来灵活调整训练轮数，从而在性能与效率之间取得平衡。

4.7.2 批次大小的影响

本节系统评估了AI-VPT方法在不同批次大小（1至128）下的性能稳定性。如图5所示，所有数据集的准确率波动幅度均小于0.5%（最大波动：ImageNe在57.5%-58.0%之间，Flowers在59.2%-59.6%之间），标准差控制在0.3%以内，证明该方法对批次尺寸具有显著鲁棒性。

这种稳定性的本质源于AI-VPT的样本级自适应机制。与依赖批次统计的传统方法（无论是AdvPT还是Adversarial Training）不同，本方法为每个输入样本生成独立的提示向量，在参数更新过程

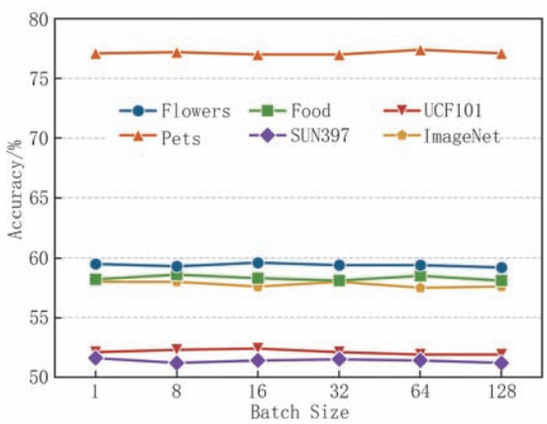

图5 不同批次大小（Batch Size）对AI-VPT的准确率（Accuracy）的影响

中完全隔离不同样本的梯度交互。这种设计使得批次尺寸的选择仅受硬件资源限制（如GPU显存或并行线程数），而不会影响模型的理论收敛行为。

与现有工作对比，本方法突破了传统参数高效微调方案对批次尺寸的敏感性限制。实验表明，从单样本推理到多样本并行，AI-VPT无需调整任何超参数即可保持性能一致，这为边缘计算设备与云计算中心的无缝协同提供了新的可能性。

4.7.3 参数初始化的影响

本小节系统评估了三种参数初始化方法（随机初始化、Xavier初始化、He初始化）对性能的影响，通过三次独立实验的准确率数据揭示不同参数初始化方法的稳定性特征。以Flowers数据集为例，图6展示的可视化设计采用横向抖动的散点图展示原始数据分布，辅以均值±标准差的误差线表征各组统

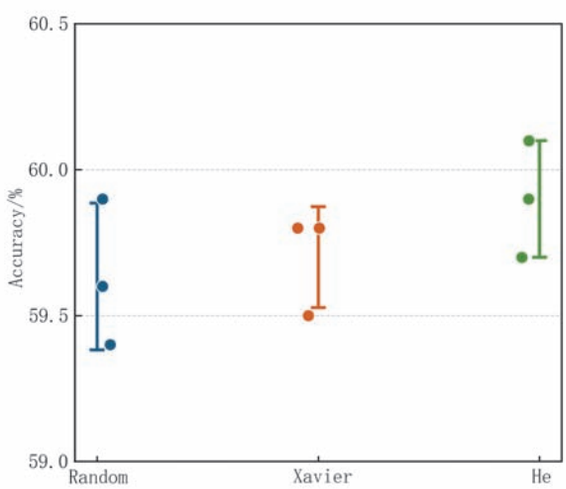

图6 不同的参数初始化方法对AI-VPT的准确率（Accuracy）的影响

计特征。数据分析表明，三种初始化方法的平均准确率分别为随机初始化(59.63±0.25)%、Xavier初始化(59.70±0.17)%、He初始化(59.90±0.20)%，组间最大差异仅为0.27%。该结果证实了AI-VPT对参数初始化的强鲁棒性，表明在实践应用中无需特别优化初始化策略即可获得稳定性能，为算法设计的初始化无关性提供了量化证据。

4.8　信息熵的有效性分析

本研究使用了信息熵作为一个衡量指标来判断该图像是否对微调有用，进而选择优化方向。从既往的研究来看，通过信息熵来判断不同视图的有效性自从首次在提出以后产生了广泛的影响，它指出对于一个本身具备强大识别能力的模型，在测试阶段进行基于信息熵的多视图筛选可以有效性减少数据中无关信息的干扰[28]。

为了验证这种策略是否有效，本部分进行了一个实验性分析以探究给定一个相同的图像，信息熵与模型所输出的预测正确的概率的关系。在ImageNet-1K验证集上，我们采用分层抽样方法选取1000张涵盖全部类别的图像，通过ViT-L/14模型获取每张图像的预测分布。计算结果表明，正确类别的预测概率与信息熵之间存在显著的统计相关性(Pearson相关系数r=－0.72，$p<0.001$，95%置信区间[－0.75，－0.69])，这一强负向关联在图7的散点分布中得以直观呈现：当信息熵低于1.5时，模型在90%以上的样本中给出正确预测；随着熵值上升至3.0-4.5区间，正确率呈现单调下降趋势；而当熵值超过4.5后，预测准确率趋近于随机猜测水平。该现象符合信息论中系统不确定性与决策置信度的理论关系，即当模型对某样本的预测分布越集中(低熵状态)，其正确识别该样本的概率越高。这一发现为基于熵值的样本筛选机制提供了理论支撑。

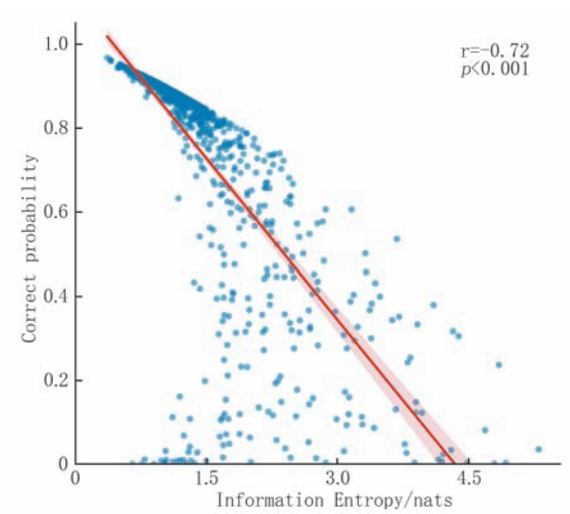

图7　预测正确概率(Correct Probability)与信息熵(Information Entropy)的相关性

5　总　结

本文提出了AI-VPT，一种基于推理阶段的对抗性视觉提示调优方法，通过优化视觉模态的嵌入，有效提升了大规模预训练视觉-语言模型的对抗鲁棒性，同时显著降低了计算成本。实验结果表明，AI-VPT在多个数据集上表现出优越的对抗防御能力，显著优于现有基线方法，尤其适用于安全性要求高的应用场景。未来研究可进一步验证AI-VPT在其他多模态模型中的适用性，结合更多增强方法，以继续提升对抗鲁棒性，并评估其在实际应用中的效果。

参　考　文　献

[1] Zhang Z K, Pang W G, Xie W J, Lyu M S, Wang Y. Review of Deep Learning for Real-Time Applications. Journal of Software, 2019, 31(9): 2654-2677(in Chinese)
(张政馗，庞为光，谢文静，吕鸣松，王义. 面向实时应用的深度学习研究综述. 软件学报，31(9)，2654-2677. 2019)

[2] Jia C, Yang Y, Xia Y, Chen Y, Parekh Z, Pham H, Le Q V, Duerig T, Song Y. Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision// Proceedings of the 38th International Conference on Machine Learning (ICML). Virtual, 2021: 4904-4916

[3] Li J, Li D, Xiong C, Hoi S. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation//Proceedings of the 39th International Conference on Machine Learning (ICML). Baltimore, USA, 2022: 12888-12900

[4] Hossain M. Z., Sohel F., Shiratuddin M. F., & Laga H. A comprehensive survey of deep learning for image captioning. ACM Computing Surveys (CsUR), 51(6), 2019: 1-36

[5] Goyal Y, Khot T, Summers-Stay D, BatraDhruv, ParikhDevi. Making the v in vqa matter: Elevating the role of image understanding in visual question answering//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA, 2017: 6904-6913

[6] Koh J Y, Fried D, Salakhutdinov R R. Generating images with multimodal language models//Advances in Neural Information Processing Systems (NeurIPS). New Orleans, USA, 2023: 21487-21506

[7] Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I, Fergus R. Intriguing Properties of Neural Networks// Proceedings of the International Conference on

Learning Representations (ICLR). Banff, Canada, 2014

[8] Shah S A, Kolouri S, Pope P, Janoyan N, Hoff W, Rohde G K. An Adversarial Approach for Explaining the Predictions of Deep Neural Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(11): 3784-3799

[9] Goodfellow I J, Shlens J, Szegedy C. Explaining and Harnessing Adversarial Examples// Proceedings of the International Conference on Learning Representations (ICLR). San Diego, USA, 2015

[10] Kurakin A, Goodfellow I, Bengio S. Adversarial Examples in the Physical World. In: Yampolskiy R V, Fox J (eds) Artificial Intelligence Safety and Security. Chapman and Hall/CRC, 2018: 99-112

[11] Madry A, Makelov A, Schmidt L, Tsipras D, Vladu A. Towards Deep Learning Models Resistant to Adversarial Attacks// Proceedings of the International Conference on Learning Representations (ICLR). Vancouver, Canada, 2018

[12] Zhang J, Ma X, Wang X, Qiu L, Wang J, Jiang YG, Sang J. Adversarial Prompt Tuning for Vision-Language Models// Proceedings of the European Conference on Computer Vision (ECCV). Milan, Italy, 2024: 56-72

[13] Carlini N, Wagner D. Towards Evaluating the Robustness of Neural Networks// Proceedings of the IEEE Symposium on Security and Privacy (SP). San Jose, USA, 2017: 39-57

[14] Moosavi-Dezfooli S M, Fawzi A, Frossard P. DeepFool: A Simple and Accurate Method to Fool Deep Neural Networks// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA, 2016: 2574-2582

[15] Papernot N, McDaniel P, Goodfellow I. Transferability in Machine Learning: from Phenomena to Black-Box Attacks using Adversarial Samples. arXiv preprint arXiv: 1605.07277, 2016

[16] Ilyas A, Engstrom L, Athalye A, Lin J. Black-box Adversarial Attacks with Limited Queries and Information//Proceedings of the 35th International Conference on Machine Learning (ICML). Stockholm, Sweden, 2018: 2137-2146

[17] Zhang C, Benz P, Lin C, et al. A survey on universal adversarial attack. arXiv preprint arXiv:2103.01498, 2021

[18] Baluja S, Fischer I. Learning to attack: Adversarial transformation networks//Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). New Orleans, USA, 2018: 2687-2695

[19] Chen P Y, Zhang H, Sharma Y, Yi J, Hsieh C J. ZOO: Zeroth Order Optimization Based Black-box Attacks to Deep Neural Networks without Training Substitute Models//Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security. Dallas, USA, 2017: 15-26

[20] Brendel W, Rauber J, Bethge M. Decision-Based Adversarial Attacks: Reliable Attacks Against Black-Box Machine Learning Models// Proceedings of the International Conference on Learning Representations (ICLR). Vancouver, Canada, 2018

[21] Cheng S, Dong Y, Pang T, et al. Improving black-box adversarial attacks with a transfer-based prior. arXiv preprint arXiv:1906.06919, 2019

[22] Zhang H, Yu Y, Jiao J, Xing E, Ghaoui L E, Jordan M. Theoretically Principled Trade-off between Robustness and Accuracy// Proceedings of the International Conference on Machine Learning (ICML). Long Beach, USA, 2019: 7472-7482

[23] Wang Y, Mao X, Yi J. Improving Adversarial Robustness Requires Revisiting Misclassified Examples// Proceedings of the International Conference on Learning Representations (ICLR). Virtual, 2020

[24] Radford A, Kim J W, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, Krueger G, Sutskever I. Learning Transferable Visual Models from Natural Language Supervision//Proceedings of the 38th International Conference on Machine Learning (ICML). Virtual, 2021: 8748-8763

[25] Zhang P, Li X, Hu X, et al. Vinvl: Revisiting visual representations in vision-language models//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Virtual, 2021: 5579-5588

[26] Lester B, Al-Rfou R, Constant N. The Power of Scale for Parameter-Efficient Prompt Tuning//Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). Punta Cana, Dominican Republic, Online, 2021: 3045-3059

[27] Zhou K, Yang J, Loy C C, Liu Z. Learning to Prompt for Vision-Language Models. International Journal of Computer Vision, 2022, 130(9): 2337-2348

[28] Shu M, Nie W, Huang D A, et al. Test-time prompt tuning for zero-shot generalization in vision-language models//Advances in Neural Information Processing Systems (NeurIPS), New Orleans, USA, 2022: 14274-14289

[29] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J, Houlsby N. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale// Proceedings of the International Conference on Learning Representations (ICLR). Virtual, 2021

[30] Nilsback M E, Zisserman A. Automated Flower Classification over a Large Number of Classes//Proceedings of the Indian Conference on Computer Vision, Graphics and Image Processing. Bhubaneswar, India, 2008: 722-729

[31] Parkhi O M, Vedaldi A, Zisserman A, Jawahar C V. Cats and Dogs// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Providence, USA, 2012: 3498-3505

[32] Bossard L, Guillaumin M, Van Gool L. Food-101-Mining Discriminative Components with Random Forests// Proceedings of the European Conference on Computer Vision (ECCV). Zurich, Switzerland, 2014: 446-461

[33] Xiao J, Hays J, Ehinger K A, Oliva A, Torralba A. SUN Database: Large-scale Scene Recognition from Abbey to Zoo// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). San Francisco, USA, 2010:

3485-3492

[34] Soomro K, Zamir A R, Shah M. UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild. arXiv preprint arXiv:1212.0402, 2012

[35] Deng J, Dong W, Socher R, Li L J, Li K, Fei-Fei L. ImageNet: A Large-Scale Hierarchical Image Database// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Miami Beach, USA, 2009: 248-255

[36] Plummer B A, Wang L, Cervantes C M, Caicedo J C, Hockenmaier J, Lazebnik, S. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models//Proceedings of the IEEE International Conference on Computer Vision (ICCV). Santiago, Chile, 2015: 2641-2649

[37] Wong E, Rice L, Kolter J Z. Fast is better than free: Revisiting adversarial training//Proceedings of the International Conference on Learning Representations (ICLR). Virtual, 2020

[38] Zhu R, Zhang Z, Liang S, Liu Z, Xu C. Learning to Transform Dynamically for Better Adversarial Transferability//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA, 2024: 24273-24283

ZHANG Jia-Ming, Ph. D. His research interests include multimedia analysis, computer vision, and vision-language models

SANG Ji-Tao, Ph. D., professor, Ph. D. supervisor. His research interests include multimedia analysis, data mining and deep learning.

YU Jian, Ph. D., professor, Ph. D. supervisor. His research interests include machine learning, data mining and deep learning.

Background

This study addresses the issue of adversarial vulnerability in large-scale pre-trained vision-language models (VLMs), a significant concern in the field of deep learning and computer vision. While VLMs have shown remarkable performance, they remain highly susceptible to adversarial noise. This imperceptible perturbation can lead to misclassification, posing substantial security risks, particularly in applications involving sensitive information.

Currently, international research has focused on improving adversarial robustness through methods like adversarial training. However, adversarial training is computationally expensive and challenging to apply to large-scale VLMs with billions of parameters. A recent approach, adversarial prompt tuning, has shown promise but only focuses on the text modality, leaving the visual modality underexplored. This paper introduces a novel method called Adversarial Inference-time Visual Prompt Tuning (AI-VPT), which enhances the adversarial robustness of VLMs by focusing on visual prompt tuning during the inference phase, without additional training overhead. AI-VPT aligns visual embeddings with adversarial image inputs to reduce the impact of adversarial noise, thus strengthening the model's defense.

The project's significance lies in improving the reliability of AI systems in real-world applications. The research team has previously contributed to the field with papers published at top conferences, including CVPR, AAAI, ACM MM and ECCV, underscoring their ongoing commitment to advancing VLM robustness.