引用格式：王卓，方鸣骐，于灵云，等. 基于概念提示微调的生成图像检测[J]. 信息对抗技术，2025，4(5)：54-65.[WANG Zhuo，FANG Mingqi，YU Lingyun，et al. Generated image detection based on conceptual prompt-tuning[J]. Information Countermeasure Technology，2025，4(5)：54-65.(in Chinese)]

基于概念提示微调的生成图像检测

王　卓，方鸣骐，于灵云*，谢洪涛

（中国科学技术大学信息科学技术学院，安徽合肥 230026）

摘　要　视觉生成人工智能技术的飞速发展，有力推动了艺术创作、医疗影像生成等领域的创新。然而，其高度逼真的生成特性也给虚假信息传播、隐私侵犯等带来安全挑战，因此急需高效的检测技术。为了解决目前生成图像检测方法在未见数据分布上泛化性不足，以及对视觉语言模型文本语义潜力利用不充分等问题，基于对比语言-图像预训练（contrastive language-image pre-training，CLIP），提出一种基于概念提示微调的生成图像检测方法。该方法通过数据驱动的显著概念提取，挖掘生成图像与真实图像的共性分布特性，生成语义化提示向量，为 CLIP 文本编码器注入丰富先验知识。基于提示微调工作和提示集成策略优化提示向量，在兼顾计算效率与预训练知识保留的同时，增强了跨模型与跨数据集的检测能力。实验结果表明，所提方法可以显著且一致地提高生成图像检测性能，在未见域上的平均准确率和精度分别提升了 5.96%和 6.37%；同时，对于常见后处理操作也具备较好的鲁棒性。消融实验进一步证明了方法的有效性与先进性，显示出其在实际应用中的潜力与可靠性。

关键词　生成图像检测；泛化性；CLIP；概念提示微调

中图分类号　TP 399　　**文章编号**　2097-163X(2025)05-0054-12

文献标志码　A　　**DOI**　10.12399/j.issn.2097-163x.2025.05.004

Generated image detection based on conceptual prompt-tuning

WANG Zhuo，FANG Mingqi，YU Lingyun*，XIE Hongtao

（Department of Electronic Engineering and Information Science，University of Science and Technology of China，Hefei 230026，China）

Abstract　The rapid development of visual generative artificial intelligence（AI）technology has strongly driven innovation in fields such as artistic creation and medical image generation. However，its highly realistic generation characteristics also pose security challenges，including the spread of disinformation and privacy violations—thus，there is an urgent need for efficient detection technologies. To address the current issues of generative image detection methods，such as insufficient generalization on unseen data distributions and inadequate utilization of the text semantic potential of vision-language models，this study proposed a generated image detection method based on conceptual prompt-tuning，leveraging contrastive language-image pre-training（CLIP）. This method extracts prominent concepts in a data-driven manner to explore the common distributional features between generated images and real images，

收稿日期：2025-07-07　　**修回日期**：2025-09-06

通信作者：于灵云，E-mail：yuly@ustc.edu.cn

基金项目：国家自然科学基金资助项目（62425114，U23B2028，62121002，62472395）

and generates semantic prompt vectors to inject rich prior knowledge into the CLIP text encoder. By optimizing the prompt vectors through prompt-tuning and a prompt ensembling strategy, it balances computational efficiency and the retention of pre-trained knowledge while enhancing detection capabilities across different models and datasets. Experimental results show that the proposed method significantly and consistently improves the performance of generated image detection, with average accuracy and precision on unseen domains increased by 5.96% and 6.37%, respectively. Additionally, it exhibits good robustness against common post-processing. Ablation experiments further verify the effectiveness and advancement of the proposed method, demonstrating its potential and reliability in practical applications.

Keywords　generated image detection; generalization; CLIP; conceptual prompt-tuning

0　引言

随着人工智能技术的迅猛发展，生成图像技术以其惊人的发展速度重塑了人工智能领域的应用格局。从生成对抗网络（generative adversarial network，GAN）[1]的开创性突破，到扩散模型（diffusion model，DM）[2]的广泛应用，生成技术已能够创作出高度逼真的图像，涵盖艺术创作、虚拟现实、医疗影像生成等多个领域，为社会创新提供了难以想象的可能。然而，技术的进步也伴随着严峻的挑战：逼真的伪造人脸图像可能被用于欺诈或恶意宣传，而高度真实的场景生成图像可能误导公众认知。人眼区分生成图像和真实图像的难度日益增加，这引起了社会的极大关注，生成图像检测技术成为网络空间内容安全与治理领域的核心课题。

传统的检测方法主要分为通过空间域或频域特征来进行模型的训练。基于空间域特征训练的方法通过像素级特征检测生成伪影，如纹理不规则性和边缘失真。早期研究利用生成模型的“数字指纹”，通过给定由目标模型生成的几百个图像，使用简单的去噪器[3-4]或基于深度学习的更复杂的方法[5-6]提取其噪声残差的平均值实现检测。后续研究进一步聚集局部补丁[7]或融合全局空间信息与局部特征[8]。在 GRAGNANIELLO 等[9]的工作中，为了保留细微的高频取证痕迹，选择不在网络的第 1 层进行任何下采样。出于同样目的，TAN 等[10]使用噪声残差而非原始图像，借助预先训练的卷积神经网络（convolutional neural network，CNN）提取梯度。这些方法在捕获局部异常和像素突变方面表现出色，但面对新型生成模型或高质量生成图像时，检测性能可能下降，即泛化能力有限，且对复杂场景和后处理的适应性较差。基于频域特征训练的方法，通过将图像数据从空间域变换到频谱域，检查在频谱域中生成图像与真实图像表现出的不同特性，有效区分二者。GAN 图像伪影在频谱域中更容易被发现，且在人工指纹谱中清晰可见[11-13]；同样，观察到 DM 图像和真实图像的径向和角光谱分布也有明显的差异[14-15]。近年来，一些学者[16-17]通过一系列微小的架构调整，复制并优化了已知生成器的生成过程，以学习处理分布外（out-of-distribution，OOD）的测试图像。此类方法特别适合检测高质量生成图像或经过后处理的图像，但随之而来的问题是，其针对高分辨率图像，计算复杂度较高，且在含复杂纹理或高噪声的图像中，频域特征容易被自然图像的频率分量掩盖，导致误检或漏检。总而言之，传统检测方法在面对训练过程中未接触的数据分布时，泛化能力存在欠缺。

随着多模态视觉语言模型的出现，众多研究人员开始利用其在大规模图像-文本数据集上训练得到的丰富先验知识，通过对齐视觉和文本特征的方式来检测生成的图像，取得了优异的检测性能。KEITA 等[18]提出了 Bi-LORA，其通过在视觉语言模型中利用双低秩自适应机制，可有效捕捉 AI 生成图像的独特特征和伪影。由于集成了视觉和文本信息，其区分真实的和生成内容的能力得到增强。RADFORD 等[19]提出了对比语言-图像预训练（contrastive language-image pre-training，CLIP），这是一种大规模模型，常用于迁移学习设置，具有新颖的泛化性，能够从众多图像中区分出真实图像和生成图像。基于此，陆续有研究人员利用其潜力进行生成图像检测。2023 年，OJHA 等[20]使用 CLIP 作为特征提取模型，并在其基础上训练线性分类头用于深度伪造检测，该方法在未接受真实图像和虚假图像分类的特定训练情况下，展现了出色的检测泛化性能。同年，WU 等[21]专注于设计文本标签，通过图像-

文本对比学习来指导 CLIP 视觉模型；2024 年，LIU 等[22]利用频率分析，并使用文本编码器作为冻结 CLIP 视觉模型的适配器来增强检测性能。尽管取得了上述进展，但已有研究强调，通过线性分类调整 CLIP 不会利用其文本组件，并且仅依赖于视觉特征，这可能导致次优性能[23-24]，生成图像检测性能还有待进一步加强。

基于以上分析，本文提出了一种基于概念提示微调的生成图像检测框架，以解决当前生成图像检测方法在未见数据分布上泛化性不足以及对视觉语言模型文本语义潜力利用不充分等问题。该方法通过提取显著概念词并转换为可学习文本提示，增强 CLIP 视觉与文本特征的跨模态对齐，借助提示微调和提示集成融合多概念语义信息，实现对生成图像的概念级特征的捕捉，进而提升对未见数据分布及复杂场景的检测性能。文中设计了一系列实验，验证了所提方法的有效性，并对结果进行了综合分析与讨论。本研究不仅丰富了生成图像检测的理论框架，也为实际应用中的安全防护提供了技术支持，对于保障网络空间内容安全与治理领域具有重要意义。

本文的主要贡献如下：

1）基于 CLIP 提出了一种基于概念提示微调的生成图像检测框架，显著提升了生成图像检测在未见数据分布上的泛化能力；

2）提出基于数据驱动的显著概念抽取并采取提示微调与集成策略，充分利用了 CLIP 的文本语义能力，弥补了现有方法在未见数据分布上的不足；

3）实验表明，本文方法可以显著且一致地提高生成图像检测性能，在未见域上的平均准确率和精度分别提升了 5.96%和 6.37%，同时对于常见后处理操作也具备较好的鲁棒性。

1 问题分析

近期研究发现，大型预训练模型（如 CLIP）与线性分类器结合，可有效提升生成图像检测技术的泛化性能[20]。然而，在相关文献中仅使用 CLIP 特征通过线性分类器就能实现优异的泛化性生成图像检测的原因并未得到应有的说明，因而在提出本文方法前，首先需要探究 CLIP 检测能力的潜在机制。

首先，本文通过使用 ClipCap[25]将 CLIP 特征解码为文本来研究该问题。ClipCap 是一种基于 Mapping Network 的 Encoder-Decoder 模型，能够根据图片生成对应的文本描述。具体而言，本文将 OJHA 等[20]工作中的线性分类器 f_c 参数线性加权变换后的图像特征定义为检测特征 v_j'，这里的检测特征 v_j' 为图像特征 v_j 和线性分类器 f_c 中权重 w 和偏置 b 参数的组合：$v_j' = v_j \cdot w + b$。为了深入了解 CLIP 如何执行检测，本文将图像特征和检测特征分别解码为文本。如图 1 所示，通过 CLIP 视觉模型处理小猫图像以提取图像特征，然后通过 ClipCap 将其解码为文本：“A picture of a very fluffy white cat”，即“一张毛茸茸的白猫的照片”。然而，当应用线性分类器来获取检测特征并将其解码为文本时，结果是：“A picture of a street with a lot of traffic”，即“一张交通繁忙的街道的照片”。值得注意的是，解码后的文本与原始图像内容没有直接关联。这一现象表明，语义信息，比如图像中存在的猫，在检测特征中不再会被获取到。所以，本文推断 CLIP 不具有固定的“真”或者“假”的语义；相反，CLIP 通过识别和匹配相似的概念来进行生成图像检测。

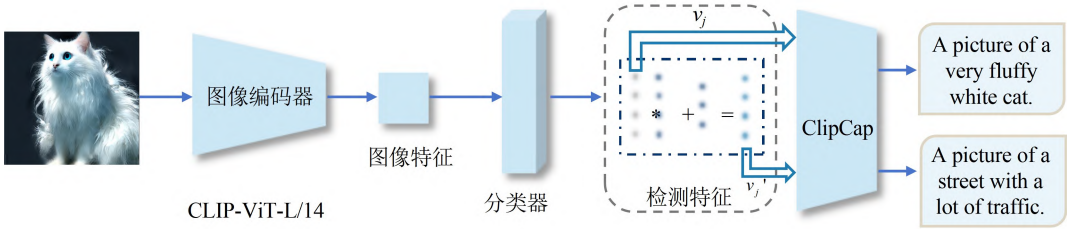


图 1 基于 ClipCap 的 CLIP 检测机制探究

Fig. 1 Exploration of CLIP detection mechanism based on ClipCap

为了验证上述假设，本文继续使用 t-SNE 特征降维对训练集的检测特征进行可视化，这里抽取了训练数据集中 3 个类别的子集应用 K-means 聚类来识别 3 个聚类中心，如图 2 所示。可以看到，蓝、黄、绿分别代表 3 个不同类别的子集，其中每个颜色中较深的为真实数据，较浅的为生成数据，两者之间形成了明显的上级分离。解码每个子集内 3 个聚类中心的检测特征的文本表示，不

难发现，解码后的文本与原始图像内容同样没有直接关联。这进一步支持了本文的假设，即 CLIP 是通过识别相似的概念来进行生成图像检测，而不是通过识别真假图像的语义。

通过将 CLIP 的检测特征解码为文本并进行 t-SNE 分析，深入探究 CLIP 检测能力的潜在机制，发现 CLIP 是通过识别相似的概念来检测生成图像。基于上述见解，本文通过词频统计，得到显著概念词，将其作为可学习的向量注入文本提示的上下文中，保持 CLIP 的视觉编码器和文本编码器冻结，使整个预训练参数保持不变，在微调过程中通过最小化分类损失进行上下文优化，进而实现 CLIP 视觉和文本能力的结合，同时结合提示集成策略，提升生成图像检测的性能。

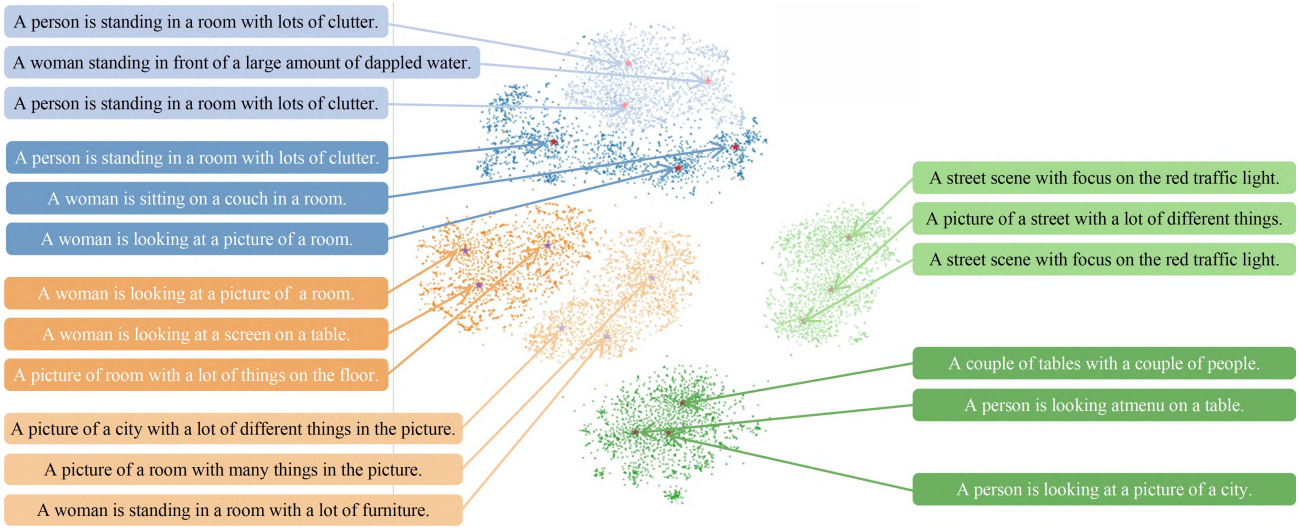


图 2　基于 t-SNE 的检测特征聚类分析

Fig. 2　Detection feature clustering analysis based on t-SNE

2　方法设计

本文提出的基于概念提示微调的生成图像检测方法的流程如图 3 所示。该方法主要由显著概念抽取、提示微调和提示集成 3 个模块组成。显著概念抽取通过对训练集的检测特征解码所得的文本进行词频分析与统计，抽取得到显著概念词；提示微调将得到的显著概念词转换为词向量，并注入可学习上下文中进行优化，形成提示；提示集成将多个显著概念词转换为对应词向量，按预设规定注入每个类别的可学习上下文中进行分别优化组合并取平均后，作为每个类别的最终提示。最后，通过训练所得模型来进行生成图像检测。

2.1　显著概念抽取

CLIP 特征通过线性分类进行生成图像检测的能力颇具研究价值。如前文分析，本文将这种能力归因于 CLIP 对相似概念的探寻机制。基于该认识，本文从训练集中真实数据集和生成数据集中抽取出能表征其检测特征的显著概念词，将其注入可学习上下文中进行优化，作为提示，使得每个类别的文本特征在特征空间中更接近对应类别的图像检测特征，从而显著提高检测性能。为此，显著概念词被设计用于增强与图像检测特征相关的词元，然后将这些词元嵌入到可学习的上下文中。

给定一个训练数据集 X，其中包含真实图像和生成图像，定义为：

$$X=\{x_j,y_j\}_{j=1}^{N},y\in\{0,1\} \tag{1}$$

式中，$y=1$ 表示图像是生成的，$y=0$ 表示图像是真实的。对于训练集中的每一幅图像 x_j，利用 ViT-L/14 提取对应的图像特征 v_j，经过 OJHA 等[20]工作中的线性分类器 f_c 的参数权重 w 和偏置 b，对图像特征进行线性加权，得到各自对应的检测特征 v_j'，定义为：

$$v_j=E_{\text{encoder,img}}(x_j) \tag{2}$$

$$v_j'=v_j\cdot w+b \tag{3}$$

对于得到的每一个检测特征 v_j'，利用 ClipCap 模型得到其对应的文本表示。整个训练数据集的检测特征文本集记为 C，定义为：

$$C=\{c_j,y_j\}_{j=1}^{N},y\in\{0,1\} \tag{4}$$

式中，c_j 表示检测特征文本。

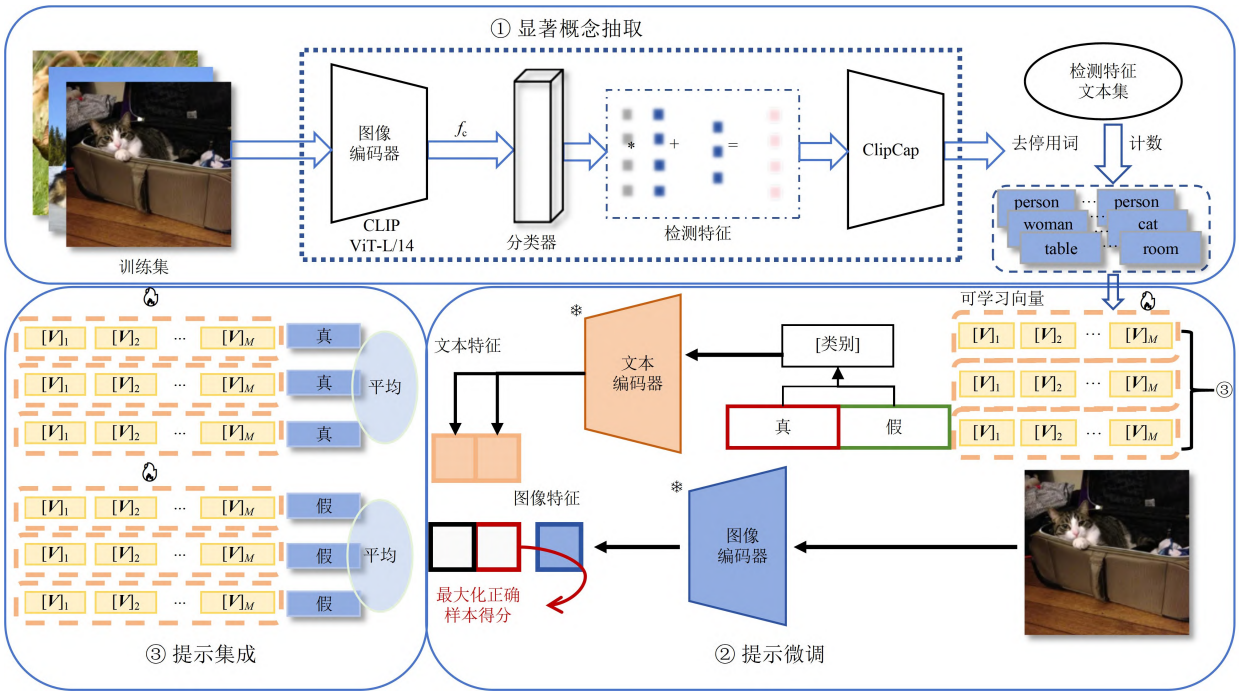


图 3 基于概念提示微调的生成图像检测方法

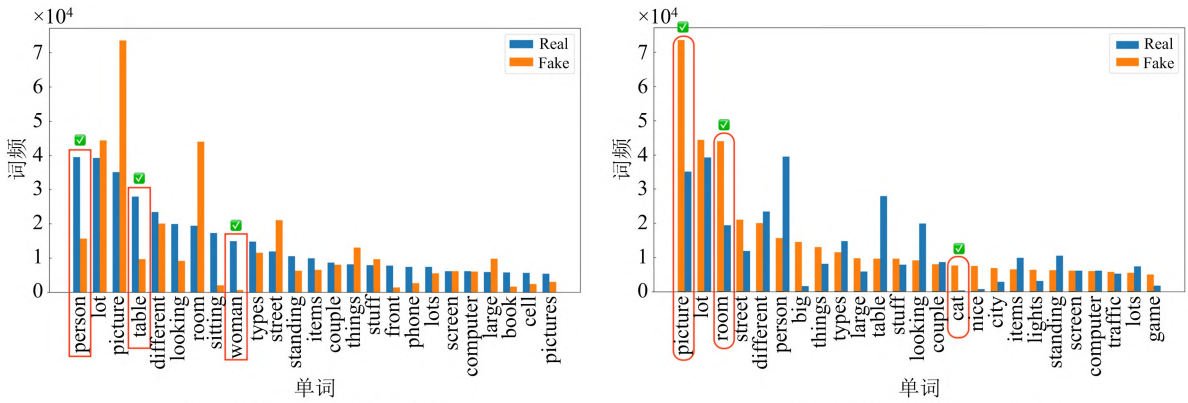
Fig. 3 The method of generated image detection based on conceptual prompt-tuning

将检测特征文本集 C 中 $y=0$ 和 $y=1$ 的文本分别抽取出来得到 C_0 和 C_1，表示真实集检测特征文本集和生成集检测特征文本集，接着分别从 C_0 和 C_1 中识别出高频出现的单词，并将其作为显著概念。具体而言，首先，对 C_0 和 C_1 中的文本数据分别进行预处理，包括将文本转换为小写、去除标点符号和停用词，以消除噪声并提高分析的准确性。随后，利用自然语言处理技术对预处理后的文本进行分词，并统计每个单词 w 的出现频率 $f_0(w)$ 和 $f_1(w)$。最后，根据词频统计结果，提取出频率最高的前若干个单词，这些高频词被认为是文本数据中的显著概念。显著概念词的筛选过程定义如下：

$$\begin{cases} W_0 = \{w \mid w \in T_{\text{top}}^K(C_0), \dfrac{f_0(w)}{f_1(w)} \geqslant \theta\} \\ W_1 = \{w \mid w \in T_{\text{top}}^K(C_1), \dfrac{f_1(w)}{f_0(w)} \geqslant \theta\} \end{cases} \tag{5}$$

式中，$T_{\text{top}}^K(C)$ 表示从 C 中提取词频最高的 K 个单词（实验中设置为 25）；θ 是一个参数（实验中设置为 2），用于确保显著概念在对应类别中具有显著的区分性；W_0 和 W_1 分别表示真实图像和生成图像的显著概念词集合。

从训练数据集中经过 ClipCap 模型和词频统计分析后得到的显著概念前 25 词的结果如图 4 所示，其中，图 4(a)为真实集结果，图 4(b)为生成集结果。



(a) 真实集显著概念Top 25词　　(b) 生成集显著概念Top 25词

图 4 训练数据集显著概念 Top 25 词

Fig. 4 Top 25 significant concepts in the training dataset

从图 4 展示的真实集和生成集显著概念 Top 25 词的结果可知，二者之间存在重叠词，如“lot”等，故为了进一步区分真实集和生成集的检测特征的显著概念，本文在选取各自的显著概念词时，将不同类别下同一词频差距 2 倍以上且仍在对应集合显著概念前 25 词中的词元作为最终筛选得到的显著概念词。（例如，真实集中“person”的“real”类对应词频是“fake”类对应词频的 2 倍及以上）。

2.2　提示微调

提示微调（prompt-tuning）最初被引入自然语言处理领域，是计算机视觉领域采用的一种相对较新的迁移学习策略。该方法通过在训练过程中学习随机初始化的提示（文本、视觉或文本和视觉结合的形式），来微调类似于 CLIP 的预训练模型。提示微调的首要目标是通过优化提示，使模型适配特定的下游任务，从而更好地与目标保持一致。

ZHOU 等[24]的开创性研究引入了上下文优化（context optimization，CoOp）来提高 CLIP 在图像分类任务中的效率。与传统的提示模板不同，CoOp 借助从数据中端到端学习的连续向量来建模上下文单词，同时冻结预训练的参数，从而避免手动提示微调；通过学习提示嵌入的方式，CoOp 对下游数据集样本的依赖最小化。本文使用 CoOp 来微调 CLIP 用于生成图像检测任务。

CoOp 在提示的上下文单词的基础上追加可学习的向量，此处的上下文单词指给定数据集的分类标签，在本文中为“real”和“fake”。这些可学习向量既可以用随机值初始化，也可以用预训练的词嵌入，本文使用的是预训练的词嵌入，并且针对不同上下文单词，使用的预训练的词嵌入也不同。具体地，给文本编码器的提示设计为如下形式：

$$\begin{cases} t=[\boldsymbol{V}]_1[\boldsymbol{V}]_2\cdots[\boldsymbol{V}]_M[\mathrm{CLASS}] \\ [\boldsymbol{V}]_M=\begin{cases} E_{\mathrm{encoder,text}}(w_0),\text{if } [\mathrm{CLASS}]=\mathrm{real} \\ E_{\mathrm{encoder,text}}(w_1),\text{if } [\mathrm{CLASS}]=\mathrm{fake} \end{cases} \end{cases} \tag{6}$$

式中，每个$[\boldsymbol{V}]_m$（$m\in\{1,\cdots,M\}$）是与词嵌入具有相同维度的向量，以 CLIP（ViT-Large）为例，维度为 768；M 是指定上下文标记数的超参数。[CLASS]表示数据集的类标记，比如在本文中，[CLASS]分别为“real”和“fake”。将提示 t 输入到文本编码器，即可得到一个表示视觉概念的分类权重向量，这里预测概率的计算公式为：

$$p(y=i\mid x_j)=\frac{\exp(\cos(E_{\mathrm{encoder,text}}(t_i),v_j)/\tau)}{\sum_{k=1}^{K}\exp(\cos(E_{\mathrm{encoder,text}}(t_k),v_j)/\tau)} \tag{7}$$

式中，K 表示类别的个数，τ 是由 CLIP 学习得到的温度参数，$\cos(\cdot,\cdot)$表示余弦相似度，每个提示 t_i 内的类标记用第 i 个类名对应的词嵌入向量替代，$i\in\{0,1\}$。此外，类标记除可如式（6）置于序列末尾外，还可置于中间或者开头，即：

$$\begin{aligned} t&=[\boldsymbol{V}]_1\cdots[\boldsymbol{V}]_{\frac{M}{2}}[\mathrm{CLASS}][\boldsymbol{V}]_{\frac{M+1}{2}}\cdots[\boldsymbol{V}]_M \\ &=[\mathrm{CLASS}][\boldsymbol{V}]_1[\boldsymbol{V}]_2\cdots[\boldsymbol{V}]_M \end{aligned} \tag{8}$$

这增加了学习的灵活性——提示既可以用补充说明填充续单元格，也可以使用终止信号提前切断句子。本文的研究结果表明，在开头添加上下文标记可以得到更好的结果，这一点在后面的实验设置部分也有提及。训练基于交叉熵最小化标准分类损失进行，梯度可以通过文本编码器一路反向传播，利用参数中编码的丰富先验知识优化上下文。连续表征的设计还支持在词嵌入空间中进行充分的探索，有利于任务相关情境的学习。交叉熵损失定义为：

$$=-\sum_{k=1}^{k}y_k\cdot\ln(p_k) \tag{9}$$

式中，y_k 表示类别 k 的真实概率，p_k 表示类别 k 的预测概率。

2.3　提示集成

提示集成（prompt ensembling）是自然语言处理领域，尤其是预训练语言模型应用中的常用技术。它通过结合多个不同的提示，提升模型的性能和鲁棒性。这里的“提示”是指用来指导模型生成或理解文本的简短文本片段或模板。提示集成的核心思想是从多个提示的输出中提取综合信息，在实际应用中，提示集成被广泛用于增强模型的表现，尤其是在需要依赖提示设计的场景下。

在本文中，提示集成被用来优化文本特征的生成过程，具体实现步骤为：首先，为每个分类类别设计多个初始上下文短语。例如，对于“fake”类别，可能会有“The picture is”“The room is”“The cat is”等短语。这些上下文短语被用来初始化一组可学习的上下文向量，这些向量会在训练过程中不断优化。对于每个上下文短语，CoOp 构造一个提示嵌入，这些提示嵌入会被输入到

CLIP 的文本编码器中，生成对应的文本特征。对于每个类别，本文方法会生成多个提示嵌入 $\{\boldsymbol{P}_1,\boldsymbol{P}_2,\cdots,\boldsymbol{P}_Z\}$，其中 Z 表示每个类别的上下文短语数量，这些嵌入对应于多个上下文短语。在前向传播中，这些提示嵌入的文本特征会被平均融合，生成该类别下一个综合的文本特征 $\boldsymbol{F}$，其计算公式为：

$$\begin{cases}\boldsymbol{P}=[\mathrm{CLASS}][\boldsymbol{V}]_1[\boldsymbol{V}]_2\cdots[\boldsymbol{V}]_M\\ \boldsymbol{F}=\dfrac{1}{Z}\displaystyle\sum_{z=1}^{Z}E_{\mathrm{encoder,text}}(\boldsymbol{P}_z)\end{cases}\tag{10}$$

同样地，每个 $[\boldsymbol{V}]_m$（$m\in\{1,\cdots,M\}$）是与词嵌入具有相同维度的向量，M 是指定上下文标记数的超参数。[CLASS]指的是数据集的类标记，在具体的训练中，该标记被添加在开头位置。那么预测概率的计算公式变为：

$$p(y=i\mid x_j)=\frac{\exp(\cos(\boldsymbol{F}_i,v_j)/\tau)}{\displaystyle\sum_{k=1}^{K}\exp(\cos(\boldsymbol{F}_k,v_j)/\tau)},i\in\{0,1\}\tag{11}$$

通过提示集成，减少了对前面步骤中单一提示的依赖。让不同提示嵌入不同的检测相关的概念语义信息，经过平均融合后生成的文本特征更全面、更稳定，能更好地在特征空间中匹配生成图像的特征，从而提高图像-文本相似度计算的准确性，最终提升分类性能。

3 实验设置及结果分析

3.1 评估指标

在性能评估上，本文使用业界公认的评价指标分类准确率（accuracy，Acc）和平均精确率（average precision，AP）来评价本方法对于生成图像内容检测泛化性以及包括鲁棒性等性能的影响，确保本方法与现有工作的可比性。具体来说，Acc 的计算公式为：

$$A_{\mathrm{Acc}}=\frac{T_{\mathrm{TP}}+T_{\mathrm{TN}}}{T_{\mathrm{TP}}+T_{\mathrm{TN}}+F_{\mathrm{FP}}+F_{\mathrm{FN}}}\tag{12}$$

式中，TP（true positive）表示真阳性，即模型正确预测为正类的正样本；TN（true negative）表示真阴性，即模型正确预测为负类的负样本；FP（false positive）表示假阳性，即模型错误地将负样本预测为正类；FN（false negative）表示假阴性，即模型错误地将正样本预测为负类。该指标能直接反映模型整体的准确率，简单高效。

AP 也是生成图像检测任务中常用的指标，特别适合评估模型在不同置信度阈值下的性能，其计算公式为：

$$\begin{cases}A_{\mathrm{AP}}=\displaystyle\sum_n(R_n-R_{n-1})P_n\\ R_{\mathrm{recall}}=\dfrac{T_{\mathrm{TP}}}{T_{\mathrm{TP}}+F_{\mathrm{FN}}}\\ P_{\mathrm{precision}}=\dfrac{T_{\mathrm{TP}}}{T_{\mathrm{TP}}+F_{\mathrm{FP}}}\end{cases}\tag{13}$$

式中，n 表示按照模型预测的置信度从高到低排序后的不同阈值点；R_n 表示第 n 个阈值点对应的召回率的值；P_n 表示第 n 个阈值点对应的精确率的值；R_{recall} 表示召回率（recall），为正类样本中被正确预测的比例；$P_{\mathrm{precision}}$ 表示精确率（precision），为预测为正类的样本中真正为正类的比例。AP 是精确率-召回率曲线（precision-recall curve）下面积的近似，通常通过对不同阈值的精确率进行加权求和计算，其综合考虑了精确率和召回率，能够反映模型在不同置信度阈值下的表现。

3.2 实验设置

实验在训练集设置上遵循 CNN-Spot[27] 和 UnivFD[20] 2 篇工作相同的协议，仅采用生成模型 ProGAN[28] 得到的 100 k 条生成数据来训练，其中对应的 100 k 条真实数据来自 LSUN 数据集，为模型训练提供了高质量的样本。在测试集上，为了使本方法的评估更加接近真实场景，本文方法在 21 个不同的数据集上进行了评估测试，主要分为基于 GAN、基于 Diffusion、商业工具[29] 和 FaceForensics++[30] 几类。训练参数设置上，和 ZHOU 等[24] 保持一致，batch_size 设置为 16，使用 RandomSampler 随机采样，图像预处理像素为 224×224；基于 OpenAI 提供的预训练 CLIP，使用 ViT-L/14 骨干网络，类别词元前置，仅优化提示中的可学习向量，冻结视觉和文本编码器，在单张 NVIDIA RTX 4090（24 GB）上训练 2 个 epoch，避免了过拟合现象。

在评估本方法有效性与优越性上，本文设计并进行了 3 个实验：

1）泛化性性能对比实验。采取的对比基线为 2 个前沿且具代表性的泛化性检测工作 CNN-Spot[27] 和 UnivFD[22]，比较基线方法与本文方法的性能指标结果；

2）鲁棒性性能测试实验。采取对测试图像

分别进行现实场景中常见的 JPEG 质量压缩（分别为 75% 和 50%）和高斯模糊 2 种不同强度（sigma（σ）值分别为 1 和 2）的后处理操作，并各自记录对应的性能指标；

3）消融实验。采取逐步移除模型训练中的不同步骤并记录对应的性能指标来验证本方法中各步骤的有效性。

3.3　结果分析

基于 3.2 节实验设置进行实验，本文方法在泛化能力和鲁棒性方面均取得了优异的效果。

3.3.1　泛化性性能对比

该实验旨在检验本文方法检测的泛化能力与其他相关前沿工作的比较。表 1 展示了泛化性性能对比实验的准确率和精准率。通过泛化性性能对比实验的结果表明，本文提出的方法在多种生成模型上均表现出优异的检测性能，尤其在 Diffusions、FaceForensics＋＋和商业工具数据分布上显著优于现有方法，整体的平均准确率达 86.69%，精准率达 96.71%，相较于对比方法，整体的平均准确率和精准率分别至少提升了 5.96%与 6.37%，最多提升了 27.37%和 24.26%。对于 FaceForensics＋＋和商业工具生成的图像，尽管优于现有方法，但检测性能相对整体较低，表明这些工具生成的图像特征复杂，仍有提升空间。总的来说，对比实验结果充分验证了本文方法通过采用提示集成与显著概念抽取相结合的策略，显著增强了在生成模型检测任务中的泛化检测能力，更加适合通用生成图像检测任务。

表 1　不同检测方法泛化性性能对比结果

Tab. 1　Comparison results of generalization performance of different detection methods

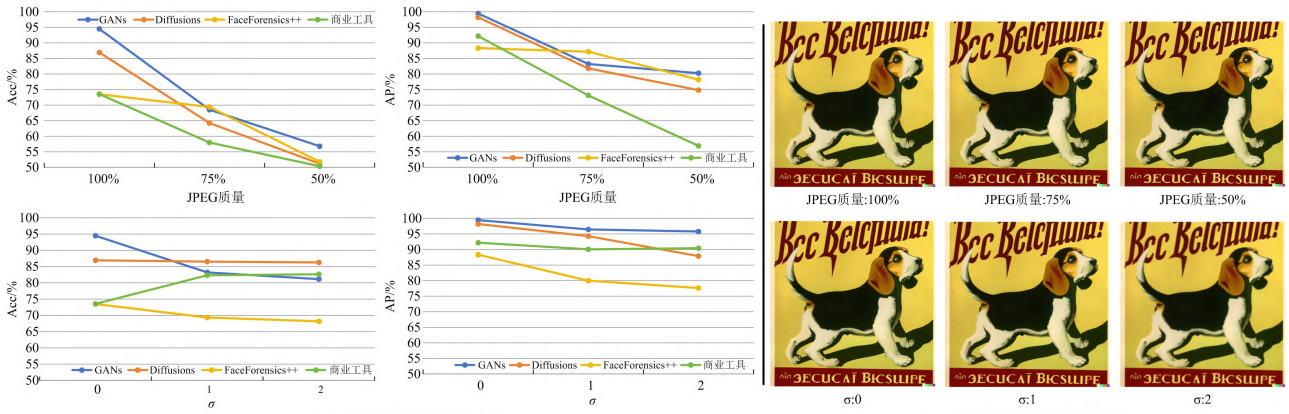
/%

模型	数据集	CNN-Spot		UnivFD	Ours
		Blur＋JPEG(50%)	Blur＋JPEG(10%)	Linear Probing	Prompt-Tuning
GANs	ProGAN	99.65/99.99	**99.99**/99.99	98.94/99.98	99.79/**99.99**
	BigGAN	58.13/82.62	67.65/83.04	94.48/98.73	93.53/**99.41**
	CycleGAN	77.80/94.70	79.50/90.09	94.20/98.91	**96.75**/**99.79**
	EG3D	50.30/55.32	72.65/95.68	57.75/79.57	**96.65**/**99.78**
	GauGAN	75.56/96.62	76.63/88.94	**94.65**/99.74	92.02/**99.84**
	StarGAN	79.99/93.88	89.72/97.17	87.49/96.06	**98.20**/**99.89**
	StyleGAN	69.80/93.25	82.10/**99.27**	85.55/95.72	**93.80**/**98.76**
	StyleGAN2	62.30/88.64	77.05/96.43	**83.40**/95.81	83.20/**98.17**
	StyleGAN3	53.42/85.33	80.68/98.63	75.42/92.20	**96.37**/**99.41**
	Taming-T	51.05/59.78	56.45/73.90	89.45/97.12	**94.60**/**99.41**
Diffusions	Glide	54.70/72.58	62.95/83.61	83.05/94.27	**97.50**/**99.79**
	Guided	52.35/65.11	62.90/83.10	79.00/92.09	**84.95**/**97.69**
	LDM	51.50/60.02	54.85/69.13	**94.35**/98.82	93.10/**99.25**
	SD	50.15/52.14	52.50/64.33	**81.89**/93.58	75.70/**96.47**
	SDXL	51.00/65.92	56.45/72.27	74.15/88.55	**83.40**/**97.83**
FaceForensics＋＋	Deepfakes	51.46/64.33	52.67/75.88	62.71/77.48	**76.20**/**92.44**
	FaceSwap	50.01/49.76	49.68/50.78	64.30/75.87	**70.79**/**84.23**
商业工具	DALL-E2	51.90/60.92	55.05/67.47	89.20/96.84	**91.95**/**99.07**
	DALL-E3	49.90/47.54	49.45/44.95	49.95/50.20	**53.25**/**79.50**
	Midjourney	50.25/51.49	51.85/57.80	61.55/76.97	**67.85**/**92.32**
	Adobe Firefly	54.55/81.53	56.00/81.97	**93.75**/**98.60**	81.05/97.87
平均 Acc/AP		59.32/72.45	65.89/79.73	80.73/90.34	**86.69**/**96.71**

3.3.2 鲁棒性性能测试

该实验旨在检验本文方法训练得到的模型在检测经过后处理的图像上的检测性能。在现实的场景中，图像在被共享在公网之前通常都不可避免地会经历后处理操作，这些操作可能会破坏图像中的伪造线索，从而显著影响检测模型的性能。根据已有研究[20,27]，本实验在维持测试集不变的情况下，对经过 2 种操作的图像进行评估：1）JPEG 压缩，测试 2 种质量级别，分别为原图质量的 75%和 50%；2）高斯模糊，测试 2 种模糊强度，其 σ 值分别为 1 和 2。

图 5 展示了鲁棒性性能测试的实验结果。折线图分别记录了在不同 JPEG 压缩级别和高斯模糊强度下，本文方法在测试集上的 Acc 值和 AP 值的变化趋势。如预期的那样，随着 JPEG 质量从 100%降至 50%以及模糊强度 σ 值从 0 增至 2，本文方法的检测性能均下降。JPEG 质量的变化对检测性能影响较大，而模糊强度的变化对检测性能影响较小。这可能是因为 JPEG 压缩破坏了图像的高频信息，削弱了伪造线索的可检测性，而高斯模糊主要影响图像平滑度，对伪造线索的破坏较轻微。但考虑到本方法没有明确针对压缩图像或模糊图像进行训练，整体的检测性能还是能维持在较优水平，这仍然是可以接受的。实验表明，本文方法通过采用提示集成与显著概念抽取相结合的策略，实现对 CLIP 视觉和文本能力结合的检测潜力的充分挖掘利用，显著增强了模型在面对常见后处理操作时的适应能力，值得一提的是本文方法对使用高斯模糊处理后的图像的鲁棒性尤为突出。



(a) JPEG压缩检测结果　　(b) 高斯模糊检测结果

图 5　基于不同后处理策略的检测方法鲁棒性结果

Fig. 5　Robustness results of detection methods based on different post-processing strategies

3.3.3 消融实验

该实验旨在验证本文方法中各步骤对提升检测性能的假设性贡献，深入剖析各步骤间的交互关系。实验保持原始测试数据集情况不变，依次测试以下仅包含特定步骤的配置：1）禁用提示集成和显著概念抽取，仅使用单一提示，提示中的可学习向量维持 ZHOU 等[24]工作原始设置，以 3 个随机值 X 初始化；2）仅使用提示集成，禁用显著概念抽取，采用多个提示，提示中的可学习向量以不同数量(3、6、9)的随机值 X 初始化；3）仅使用显著概念抽取，禁用提示集成，使用单一提示，提示中的可学习向量以显著概念词对应的预训练词嵌入初始化；4）同时使用提示集成和显著概念抽取，采用多个提示，提示中的可学习向量以多个显著概念词对应的预训练词嵌入初始化。每种配置仅记录准确率的平均值。表 2 展示了消融实验的结果。

可以看到本文方法中各步骤均对最后的整体检测效果有提升。当仅使用提示集成时，平均精准率为 82.41%，提升了 1.54%。这一提升源于提示集成通过多提示增加了输入多样性，缓解了单一提示的过拟合风险，增强了模型对未见数据的鲁棒性；但由于随机初始化的语义不足，模型学习到的提示向量可能不够语义化，限制了提升幅度。当仅使用显著概念抽取，平均精准率显著提升至 85.16%，提升了 4.29%。这一大幅提升归因于预训练词嵌入提供的语义先验，使模型从更有意义的起点开始学习，加速了收敛并增强了对任务相关特征的捕捉能力。尽管单一提示的局限性仍存在，但语义先验的引入弥补了这一

不足,显示出显著概念抽取对泛化能力的显著贡献。最后,同时使用提示集成和显著概念抽取,采用多个提示并以预训练词嵌入初始化,平均精准率为最高,达到 86.69%。这表明提示集成和显著概念抽取的结合效果最佳,多样性提示与语义先验协同作用,使模型既能从多个视角学习,又能从语义化的起点出发,进一步提升了检测能力。

表 2　消融实验结果

Tab. 2　The results of ablation study

/%

模型	数据集	包含步骤			
		/	揭示集成	显著概念抽取	提示集成 & 显著概念抽取
GAN	ProGAN	99.67	**99.83**	99.70	99.79
	BigGAN	92.70	92.42	92.87	**93.53**
	CycleGAN	96.00	96.20	96.47	**96.75**
	EG3D	85.30	86.45	92.35	**96.65**
	GauGAN	88.63	89.55	89.67	**92.02**
	StarGAN	98.97	97.77	**99.42**	98.20
	StyleGAN	91.90	93.70	93.25	**93.80**
	StyleGAN2	72.35	81.15	**83.45**	83.20
	StyleGAN3	79.26	86.84	92.78	**96.37**
	Taming-T	92.30	92.00	94.10	**94.60**
Diffusions	Glide	95.45	96.15	96.95	**97.50**
	Guided	80.10	81.80	84.55	**84.95**
	LDM	93.15	92.75	**94.35**	93.10
	SD	68.00	70.10	74.50	**75.70**
	SDXL	71.30	76.45	77.20	**83.40**
Face Forensics++	Deepfakes	69.89	70.38	75.90	**76.20**
	FaceSwap	56.64	57.28	68.83	**70.79**
商业工具	DALL-E2	93.00	90.20	**93.90**	91.95
	DALL-E3	51.00	51.95	52.10	**53.25**
	Midjourney	57.15	59.40	61.70	**67.85**
	Adobe Firefly	66.20	68.30	74.30	**81.05**
AP		80.87	82.41	85.16	**86.69**

4　结束语

本文针对当前生成图像检测方法泛化性检测能力不足以及对视觉语言模型文本语义潜力利用不充分等问题,提出了一种基于概念提示微调的生成图像检测方法。通过数据驱动的显著概念抽取,挖掘生成与真实图像的共性特征,生成语义化提示向量,为 CLIP 文本编码器注入丰富先验;基于提示微调工作和提示集成策略优化提示向量,兼顾计算效率与预训练知识保留的同时增强了跨模型与跨数据集的检测能力。

为了评估本文方法的可推广性和有效性,本文在 4 个类别下 21 个不同的数据集上进行了测试。实验结果显示,相较于基线方法,本文方法

的工作性能都有明显提升。泛化性测试中平均准确率达 86.69%，平均精准率达 96.71%，相较于基线方法分别提升了 5.96%和 6.37%。鲁棒性测试表明，本方法能有效应对 JPEG 压缩、高斯模糊等后处理操作，消融实验进一步证明本方法中每个步骤均对生成图像检测的整体工作性能起到了促进作用。因此，本文提供了一种有效的通用生成图像内容检测方法，为生成图像检测在网络空间内容安全与治理领域中的实际应用提供了坚实的理论支持与实践验证。

参 考 文 献

[1] GOODFELLOW I J, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets[C]//Proceedings of the 28th International Conference on Neural Information Processing Systems. [S. l. : s. n.], 2014:2672-2680.

[2] HO J, JAIN A, ABBEEL P. Denoising diffusion probabilistic models [C]//Proceedings of the 34th International Conference on Neural Information Processing Systems. [S. l. : s. n.], 2020:6840-6851.

[3] MARRA F, GRAGNANIELLO D, VERDOLIVA L, et al. Do GANs leave artificial fingerprints? [C]// Proceedings of 2019 IEEE Conference on Multimedia Information Processing and Retrieval. [S. l.]: IEEE, 2019: 506-511.

[4] YU N, DAVIS L, FRITZ M. Attributing fake images to GANs: learning and analyzing GAN fingerprints [C]//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. [S. l.]: IEEE, 2019: 7556-7565.

[5] LIU B, YANG F, BI X L, et al. Detecting generated images by real images[C]//Proceedings of the 17th European Conference on Computer Vision. [S. l.]: Springer, 2022: 95-110.

[6] SINITSA S, FRIED O. Deep image fingerprint: towards low budget synthetic image detector and model lineage analysis[EB/OL]. (2023-03-19)[2025-07-20]. https://arxiv.org/abs/2303.10762.

[7] CHAI L, BAU D, LIM S-N, et al. What makes fake images detectable? Understanding properties that generalize [C]//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer, 2020: 103-120.

[8] JU Y, JIA S, KE L P, et al. Fusing global and local features for generalized AI-synthesized image detection [C]//Proceedings of 2022 IEEE International Conference on Image Processing. [S. l.]: IEEE, 2022: 3465-3469.

[9] GRAGNANIELLO D, COZZOLINO D, MARRA F, et al. Are GAN generated images easy to detect? A critical analysis of the state-of-the-art[C]//Proceedings of 2021 IEEE International Conference on Multimedia and Expo. [S. l.]: IEEE, 2021:1-6.

[10] TAN C C, ZHAO Y, WEI S K, et al. Learning on gradients: generalized artifacts representation for GAN-generated images detection[C]//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S. l.]: IEEE, 2023: 12105-12114.

[11] DURALL R, KEUPER M, KEUPER J. Watch your up-convolution: CNN based generative deep neural networks are failing to reproduce spectral distributions [C]//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S. l.]: IEEE, 2020: 7890-7899.

[12] DZANIC T, SHAH K, WITHERDEN F. Fourier spectrum discrepancies in deep network generated images [C]//Proceedings of the 34th International Conference on Neural Information Processing Systems. [S. l. : s. n.], 2020:3022-3032.

[13] FRANK J, EISENHOFER T, SCHÖNHERR L, et al. Leveraging frequency analysis for deep fake image recognition[C]//Proceedings of 2020 International Conference on Machine Learning. [S. l. : s. n.], 2020: 3247-3258.

[14] CORVI R, COZZOLINO D, POGGI G, et al. Intriguing properties of synthetic images: from generative adversarial networks to diffusion models[C]//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S. l.]: IEEE, 2023: 973-982.

[15] YANG X Y, ZHOU D Q, FENG J S, et al. Diffusion probabilistic model made slim [C]//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S. l.]: IEEE, 2023: 22552-22562.

[16] JEONG Y, KIM D, RO Y, et al. Fingerprintnet: synthesized fingerprints for generated image detection [C]//Proceedings of the 17th European Conference on Computer Vision. [S. l.]: Springer, 2022: 76-94.

[17] ZHANG X, KARAMAN S, CHANG S F. Detecting and simulating artifacts in GAN fake images [C]// Proceedings of 2019 IEEE International Workshop on Information Forensics and Security. [S. l.]: IEEE, 2019: 1-6.

[18] KEITA M, HAMIDOUCHE W, EUTAMENE H B, et al. Bi-LORA: a vision-language approach for synthetic image detection[EB/OL]. (2024-04-02)[2025-07-20]. https://arxiv.org/abs/2404.01959.

[19] RADFORD A，KIM J W，HALLACY C，et al. Learning transferable visual models from natural language supervision[C]//Proceedings of 2021 International Conference on Machine Learning. [S. l.：s. n.]，2021：8748-8763.

[20] OJHA U，LI Y H，LEE Y J. Towards universal fake image detectors that generalize across generative models[C]//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S. l.]：IEEE，2023：24480-24489.

[21] WU H W，ZHOU J T，ZHANG S L. Generalizable synthetic image detection via language-guided contrastive learning[EB/OL]. (2023-05-23)[2025-07-20]. https://arxiv.org/abs/2305.13800.

[22] LIU H，TAN Z C，TAN C C，et al. Forgery-aware adaptive transformer for generalizable synthetic image detection[C]//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S. l.]：IEEE，2024：10770-10780.

[23] KIM K，LASKIN M，MORDATCH I，et al. How to adapt your large-scale vision-and-language model[EB/OL]. (2022-01-29)[2025-07-20]. https://openreview.net/forum? id=EhwEUb2ynIa.

[24] ZHOU K Y，YANG J K，LOY C C，et al. Learning to prompt for vision-language models[J]. International Journal of Computer Vision，2022，130(9)：2337-2348.

[25] MOKADY R，HERTZ A，BERMANO A H. ClipCap：CLIP prefix for image captioning[EB/OL]. (2021-11-18)[2025-07-20]. https://arxiv.org/abs/2111.09734.

[26] DOSOVITSKIY A，BEYER L，KOLESNIKOV A，et al. An image is worth 16x16 words：transformers for image recognition at scale[C]//Proceedings of the 6th International Conference on Learning Representations. [S. l.：s. n.]，2021：611-631.

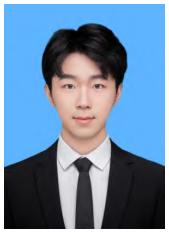
[27] WANG S Y，WANG O，ZHANG R，et al. CNN-generated images are surprisingly easy to spot... for now[C]//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S. l.]：IEEE，2020：8695-8704.

[28] KARRAS T，AILA T，LAINE S，et al. Progressive growing of GANs for improved quality，stability，and variation[EB/OL]. (2017-10-27)[2025-07-20]. https://arxiv.org/abs/1710.10196.

[29] CORVI R，COZZOLINO D，ZINGARINI G，et al. On the detection of synthetic images generated by diffusion models[C]//Proceedings of 2023 IEEE International Conference on Acoustics，Speech and Signal Processing. [S. l.]：IEEE，2023：1-5.

[30] RÖSSLE A，COZZOLINO D，VERDOLIVA L，et al. FaceForensics＋＋：learning to detect manipulated facial images[C]//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. [S. l.]：IEEE，2019：1-11.

作 者 简 介



王　卓

男，2003 年生，硕士研究生，研究方向为人工智能、深度学习、深度伪造与检测

E-mail：zhuowang70@gmail.com

方鸣骐

男，1999 年生，博士研究生，研究方向为人工智能、深度学习、深度伪造与检测

E-mail：mqfang@mail.ustc.edu.cn

于灵云

女，1992 年生，博士，副研究员，研究方向为人工智能、智能内容生成与安全

E-mail：yuly@ustc.edu.cn

谢洪涛

男，1983 年生，博士，教授，博士研究生导师，研究方向为人工智能、网络空间安全

E-mail：htxie@ustc.edu.cn

责任编辑　殷文卓