细节关注：用于高分辨率人工智能生成图像检测的全局-局部注意力

Lawrence H
anlawrence.hanyl@gmail.comGit Hub 项目代码

arXiv:2601.00141v1 [cs.CV] 1 Jan 2026

摘要

生成式人工智能的快速发展使人工智能生成图像变得愈发逼真且分辨率不断提高。大多数人工智能生成图像检测架构通常会在将图像输入模型前对其进行下采样，这可能导致细粒度细节丢失。本文提出**GLASS**（**全**球-**局**部**注**意力与**分**层**采**样），这是一种将全局调整大小视图与多个随机采样的局部裁剪块相结合的架构。这些裁剪块是原始分辨率的图像区域，通过空间分层采样高效选取，并使用基于注意力的评分进行聚合。GLASS 可集成到视觉模型中，以利用任意尺寸图像的全局与局部信息。本文采用视觉 Transformer、ResNet 和 ConvNeXt 作为骨干网络；实验表明，在可行的计算约束下，GLASS 的预测性能优于标准迁移学习。

1 引言

“这是人工智能生成的吗？”如今已逐渐成为我们在网上看到任何媒体内容时几乎脱口而出的反应。人工智能的迅速发展，使模型能够生成极其逼真、分辨率极高的图像，这些图像在人眼看来几乎与真实图像难以区分。随着这类生成式人工智能广泛普及，相关风险与担忧也同样迅速增加。

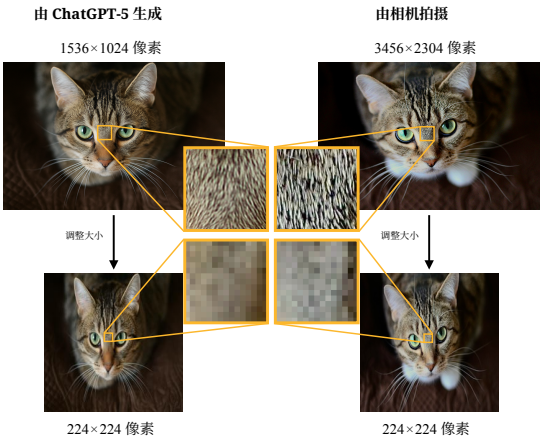

图1：**ChatGPT-5 生成图像与真实图像在调整大小前后的像素级对比**。调整大小后，细节会丢失。

许多用于人工智能生成图像检测的架构通常会对图像进行预处理，将其调整为固定分辨率，最常见的做法是 224 × 224 像素，而不考虑图像的原始尺寸（Mu 等人，2025）。这样可确保所有图像都符合基础模型的输入要求。然而，这种方法可能会无意中丢弃图像生成器生成的细粒度细节，尤其是在处理高分辨率图像时。因此，后续的分类模型可能会丢失原本有助于做出更准确判断的相关信息（Wang 等人，2020）。

本文介绍了 **GLASS**（**G**lobal-**L**ocal**A**ttention with**S**tratified**S**ampling），这是一种使模型能够同时利用图像宏观与微观信息的架构。全局调整大小后的视图使模型能够捕捉语义上下文，而局部裁剪块则让模型专注于细粒度细节。局部裁剪块在原始分辨率下随机采样，并采用空间分层，以确保在较大的非重叠面积上获得最低数量的裁剪块。通过基于注意力的加权（Bahdanau 等人，2016）机制对裁剪块进行聚合，从而为包含更多信息特征的区域分配更高的重要性。该架构使用三种骨干模型进行评估：视觉Transformer（ViT）（Dosovitsky 等人，2021）、ConvNeXt（Liu 等人，2022）和 ResNet（He 等人，2015）。其性能与仅使用全局缩放输入的标准迁移学习进行了比较。

本文的贡献如下：（**1**）引入 GLASS 架构，该架构利用分层采样和基于注意力的聚合来处理采样裁剪块。（**2**）在三种不同的骨干模型上，将 GLASS 与标准迁移学习进行比较，对性能进行了全面评估。

2 相关工作

本节回顾了利用全局与局部信息进行人工智能生成图像检测的相关研究，重点介绍局部裁剪块的使用方式以及选择局部区域的策略。

2.1 应用局部裁剪块

人工智能生成图像检测领域的研究探索了使用局部高分辨率裁剪块，以保留下采样过程中可能丢失的细微伪影。由Ju 等人（2022）提出的架构采用双分支系统，并利用多头注意力机制融合全局与局部特征。AIDE（Yan 等人，2025）使用 ResNet-50 和 CLIP-ConvNeXt 作为骨干模型，并结合全局与局部特征以及离散余弦变换评分来检测低级伪影。图块打乱（Chu 等人，2025）通过随机排列局部图块来提升泛化能力。Smash&Reconstruction（Zhong 等人，2024）将图像随机裁剪为局部图块，并仅根据这些图块的重建图像进行分类。相比之下，HiDA-Net（Mu 等人，2025）使用 ViT 骨干模型和特征聚合模块，将调整大小后的全局视图与覆盖整幅图像的原始尺寸局部裁剪块相结合，以确保使用所有像素。GLASS 同样采用带有注意力机制的双分支系统进行特征聚合，但仅对局部裁剪块使用加性注意力评分（Bahdanau 等人，2016；Ilse 等人，2018），以进行重要性加权。

2.2 局部裁剪块选择

虽然多种方法都着重使用局部裁剪块，但如何高效地选择具有信息量的裁剪块同样至关重要。PSM（Ju 等人，2022）模块通过图块评分而非随机选择，筛选出信息量最大的裁剪块。其裁剪尺寸由窗口大小决定，随后调整大小，再输入 ResNet-50。HiDA-Net（Mu 等人，2025）以原始分辨率对整幅图像进行采样，以确保局部裁剪块全面覆盖。TextureCrop（Konstantinidou 等人，2025）使用纹理选择度量选出排名前 10 的裁剪块。GLASS 采用空间分层的随机选择，以兼顾效率与稳健性。

3 方法

3.1 GLASS 概述

GLASS 架构采用双流设计（Simonyan 和 Zisserman，2014）。全局流处理调整大小后的完整图像，以捕捉高级语义内容；局部流则聚焦于原始分辨率的局部裁剪块，以检测细粒度细节。这种双流设计降低了单个模型必须区分经过全局调整大小的图像和局部裁剪图像的风险。两个独立的模型分别针对各自的任务进行设计和专门化训练，能够学习专门适用于其图像尺度和功能的权重。GLASS 的关键组成部分是局部采样器和简单注意力机制的运用。局部采样器根据图像尺寸动态提取裁剪块。注意力机制聚焦于从多个局部裁剪块中提取的特征，并将它们聚合成一个信息量丰富的特征。全局图像输入全局模型，其输出与聚合后的局部特征拼接，然后将组合后的特征嵌入传递给分类器进行分类。

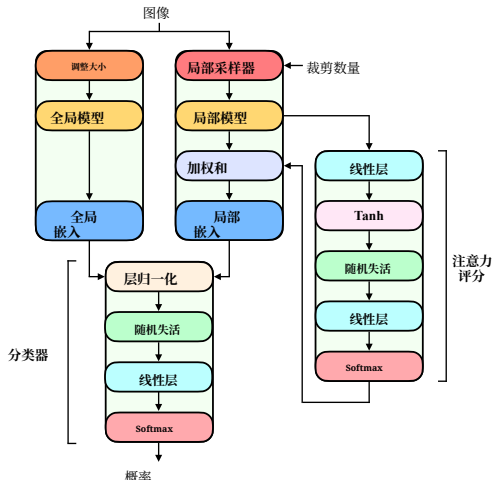

图 2：**GLASS 架构图**。

3.2 多尺度图像处理

3.2.1 全局缩放

给定一幅分辨率为（H，W 像素的自然 H，$W \geq 224$ RGB 图像 $\boldsymbol{x} \in \mathbb{R}^{3\times H\times W}$，我们使用双线性插值将其调整为 224，224 的正方形尺寸，这是预训练模型所要求的输入尺寸。

3.2.2 LocalCrop 采样器

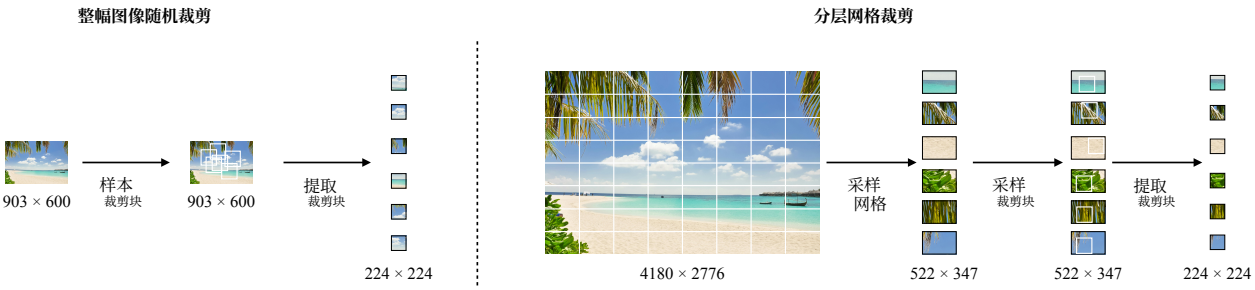

图 3：**局部裁剪采样器的两种裁剪策略**。每张图像下方标注了像素尺寸 e.

对于自然 H，$W \geq 224$，给定一张分辨率为（H，W 像素的 RGB 图像 $\boldsymbol{x} \in \mathbb{R}^{3\times H\times W}$，以及裁剪数量 $n \in \mathbb{N}$，我们定义网格尺寸为

$$G = \min\left(\left\lfloor \frac{\min(H,W)}{224} \right\rfloor, 8\right),$$

将图像划分为 G^2 个单元。常数 8 是网格单元数量的上限。这样既能确保计算效率，也能通过增大单元尺寸引入额外的空间随机性。每个单元的尺寸为（$\lfloor \frac{H}{G} \rfloor$，$\lfloor \frac{W}{G} \rfloor$），确保两个维度都至少为 224 像素。局部采样器针对两种不同情况采用两种裁剪策略。当网格单元数量少于裁剪块数量时，即 $G^2 < n$，我们在整幅图像上均匀随机采样 n 个裁剪块，不使用网格。当网格单元数量大于或等于裁剪块数量时，即 $G^2 \geq n$，我们随机选取 n 个不同的网格单元，并从每个单元中随机采样一个裁剪块。通过这两种策略得到的所有裁剪块均为 RGB 图像，分辨率为（224，224），对应于原始分辨率图像中的一个 224×224 像素方形区域。

表1： **对于不同图像尺寸和裁剪数量，预期覆盖面积的百分比为** n。

图像尺寸	$n = 2$	$n = 4$	$n = 6$	$n = 8$	$n = 10$	$n = 12$	$n = 14$	$n = 16$
224×224	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
256×256	83.9	90.0	92.8	94.4	95.4	96.2	96.7	97.1
640×480	32.7	65.3	53.8	60.5	65.3	69.0	71.9	74.2
1024×768	12.8	25.5	38.3	51.0	43.8	48.8	53.1	56.8
1280×720	10.9	21.8	32.7	43.6	39.5	44.5	48.8	52.6
1920×1080	4.8	9.7	14.5	19.4	24.2	29.0	33.9	38.7
2560×1440	2.7	5.4	8.2	10.9	13.6	16.3	19.1	21.8
3840×2160	1.2	2.4	3.6	4.8	6.0	7.3	8.5	9.7
5120×2880	0.7	1.4	2.0	2.7	3.4	4.1	4.8	5.4
7680×4320	0.3	0.6	0.9	1.2	1.5	1.8	2.1	2.4

图例： 纟 = 分层网格裁剪， =整幅图像随机裁剪。

如表 1 所示，低分辨率图像通常采用整幅图像随机裁剪，而对于高分辨率图像，主要采用分层网格裁剪策略。

3.3 骨干模型

尽管 GLASS 架构也能从头开始训练，但为提高效率并取得更好的性能，它采用预训练模型。本文使用了三种常见的骨干模型进行迁移学习，包括 ViT-Base/16（Dosovitskiy 等人，2021）、ResNet-50（He 等人，2015）以及 ConvNeXt-Tiny（Liu 等人，2022）。本文的目的并非找出最优骨干模型，而是展示 GLASS 架构在不同类型骨干模型上的有效性。

3.4 注意力机制

GLASS 使用可训练的加性注意力机制（Bahdanau 等人，2016）为采样得到的局部裁剪块分配重要性分数。该机制由一个两层全连接多层感知机组成，采用双曲正切（Ilse 等人，2018）激活函数和随机失活（Srivastava 等人，2014）进行正则化。注意力层将骨干模型为每个局部裁剪块生成的原始嵌入输出处理为重要性分数。重要性分数较高的裁剪块表明其中包含对分类决策更有价值的特征。所有局部裁剪块均按其重要性分数加权，聚合为单一表示。这样可以确保分数较高的裁剪块对分类的贡献更大。注意力层的参数通过训练学习，以最大限度地提升其识别重要特征的能力。

3.5 分类器

为获得最终预测结果，我们添加一个轻量级分类器，以全局和局部特征嵌入作为输入，并输出概率。全局嵌入是全局模型的输出，而局部嵌入则是局部模型经注意力聚合后的结果。由于两个骨干模型都已生成高级特征嵌入，使用轻量级分类器可以减少不必要的复杂度，并提升模型的泛化能力。

3.6 数据集

我们使用了一个包含 12,000 张图像的数据集，该数据集是“ai-generated-images-vs-real-images”（Zhang，2023）数据集的一个子集。该数据集由真实图像和人工智能生成图像平均构成。真实图像中四分之三来自 Pexels 和 Unsplash，剩余四分之一来自 WikiArt。人工智能生成图像则来源于 Stable Diffusion、MidJourney 和 DALL-E，三者各贡献虚假类别的三分之一。

表2：**完整数据集统计信息**。由于数据集中包含横向图像和纵向图像，平均图像尺寸接近方形宽高比。

类别	样本数	平均高度（像素）	平均宽度（像素）
real	6,000	2,205	2,212
fake	6,000	1,188	1,234

完整数据集经过随机划分，其中 8,400 张（70%）图像用于训练，1,800 张（15%）图像用于验证，另有 1,800 张图像（15%）用于测试。所有模型，包括标准迁移学习模型和基于 GLASS 的模型，均使用相同的预处理，在同一数据集上进行训练、验证和测试。

3.7 超参数调优

使用 Optuna（Akiba 等人，2019）框架对超参数进行微调。标准迁移学习模型分配了 50 次试验。基于 GLASS 的模型拥有两倍数量的超参数，因此分配了 100 次试验。每次试验最多进行 25 个训练轮次。为提高效率，使用 Hyperband 剪枝器提前停止表现不佳的试验。超参数搜索空间列于附录中的表6和表7 中。GLASS 架构中的两个骨干模型分别使用不同的学习率和权重衰减值。这样可以确保 GLASS 架构的每个组件都针对其功能进行优化。

4 结果

4.1 裁剪数量

获得能够达到最高验证准确率的超参数配置后，我们使用这些配置重新训练模型，同时仅改变裁剪数量，并将所有模型的批量大小固定为 32 张图像。对所有模型采用固定批量大小，可确保计算效率的比较公平，避免批量大小不一致影响内存使用量和训练时间。

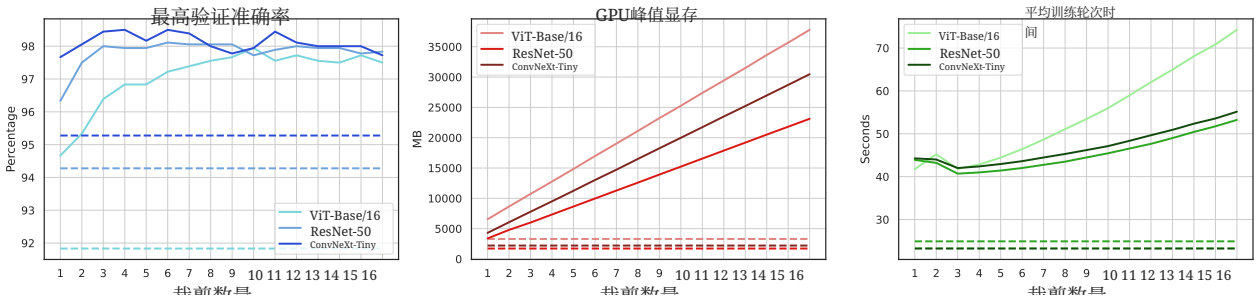

图 4：**裁剪块数量对验证准确率、内存和训练时间的影响**。 实线表示基于 GLASS 的模型的结果，虚线表示标准迁移学习的结果。

标准迁移学习模型的验证准确率、内存使用量和训练时间保持不变，因为它们只使用全局调整大小后的视图，不包含任何局部裁剪块。基于 GLASS 的模型的验证准确率会随着裁剪块数量的增加而提高，最终趋于平稳，并在达到某个阈值后出现波动。最初增加裁剪块数量可能会提供更多空间信息，从而捕捉局部细节；继续增加则会带来递减的收益。这表明，适量的裁剪块就能充分覆盖识别细粒度伪影和合成模式所需的局部特征。内存使用量和训练时间都会随着裁剪块数量的增加而逐步上升。

表 3：**内存使用量随裁剪数量线性增长 n，范围为 1 至 16**。测量 GPU 峰值显存使用量（MB），并为 GLASS 模型提供近似线性回归拟合及对应的 R^2 值。

骨干模型	线性回归	R^2	标准迁移学习
ViT-Base/16	$4463.16 + 2081.92 \times n$	1.000	3302.07
ConvNeXt-Tiny	$2545.78 + 1744.40 \times n$	1.000	2212.53
ResNet-50	$2088.69 + 1314.73 \times n$	1.000	1722.31

为分析增加裁剪块数量的影响，在图像批量大小固定为 32 的情况下，对 GPU 峰值显存进行线性分解。每种裁剪块数量对应的内存使用量，按全部 25 个训练轮次中的 GPU 峰值显存最大值进行测量。近似线性回归拟合的截距代表静态内存组成部分，而斜率表示每增加一个裁剪块带来的内存增量。三个模型的线性回归结果均显示 R^2 值为 1。000，表明 GPU 峰值显存与裁剪块数量之间存在完美的线性关系。

表 4：**裁剪块数量从 1 增加到 16 时训练时间的线性扩展** n。测量每个训练轮次的平均耗时（秒），并给出了GLASS模型对应 R^2 值的近似线性回归拟合。

骨干模型	线性回归	R^2	标准迁移学习
ViT-Base/16	$35.44 + 2.23 \times n$	0.948	23.31
ConvNeXt-Tiny	$40。05 + 0。82 \times n$	0.870	23.23
ResNet-50	$39.12 + 0.74 \times n$	0.804	24.91

训练时间也采用类似的线性分解进行分析，测量指标为全部 25 个训练轮次的平均单轮时间（秒）。近似线性回归拟合的截距代表每个训练轮次的固定时间分量，斜率则表示每增加一个裁剪块所带来的训练轮次时间增量。对三个模型进行线性回归后，平均 R^2 值为 0.874，表明平均训练轮次时间与裁剪数量之间存在很强的线性关系。

4.2 最终模型的定量性能

最终模型训练时，将训练数据集和验证数据集合并并重新打乱，共计 10,200 张图像（完整数据集的 85%）。使用各模型各自表现最佳的超参数训练最终模型。基于 GLASS 的模型采用超参数调优期间确定的原始最优批次大小，以及在第4.1节中取得最高验证准确率的裁剪数量。这些最终超参数配置见附录中的表8 和表9 。

表5：**基线模型与 GLASS 模型在测试数据集上的性能比较**。结果以不同随机种子运行 5 次的均值 ± 标准差报告，训练时使用最佳超参数。ECE 指期望校准误差。最佳性能以粗体标示。

模型	准确率（%）	AUC	F1	精确率	召回率	ECE
标准 ViT-Base/16	93.244 ± 0.499	0.972 ± 0.006	0.932 ± 0.005	0.933 ± 0.005	0.932 ± 0.005	**0.033 ± 0.003**
GLASS ViT-Base/16	**97.533 ± 0.203**	**0.997 ± 0.000**	**0.975 ± 0.002**	**0.975 ± 0.002**	**0.975 ± 0.002**	0.058 ± 0.003
标准 ConvNeXt-Tiny	96.133 ± 0.145	0.990 ± 0.002	0.961 ± 0.001	0.961 ± 0.001	0.961 ± 0.001	**0.041 ± 0.002**
GLASS ConvNeXt-Tiny	**98.611 ± 0.232**	**0.998 ± 0.001**	**0.986 ± 0.002**	**0.986 ± 0.002**	**0.986 ± 0.002**	0.048 ± 0.001
标准 ResNet-50	93.656 ± 0.674	0.981 ± 0.002	0.937 ± 0.007	0.937 ± 0.007	0.937 ± 0.007	**0.040 ± 0.005**
GLASS ResNet-50	**97.656 ± 0.499**	**0.997 ± 0.001**	**0.977 ± 0.005**	**0.977 ± 0.005**	**0.977 ± 0.005**	0.052 ± 0.004

为确保比较公平，基于 GLASS 的模型训练轮次与标准模型相同。此外，尽管搜索空间包含的超参数数量是标准模型的两倍，基于 GLASS 的模型所用计算预算也仅为其两倍，而实际上，搜索复杂度的增长通常远快于线性增长。尽管如此，与仅使用全局调整大小视图的标准迁移学习相比，基于 GLASS 的模型始终取得更高的准确率、曲线下面积、F1 分数、精确率和召回率。不过，它们的期望校准误差也始终更高。

4.3 权重分布

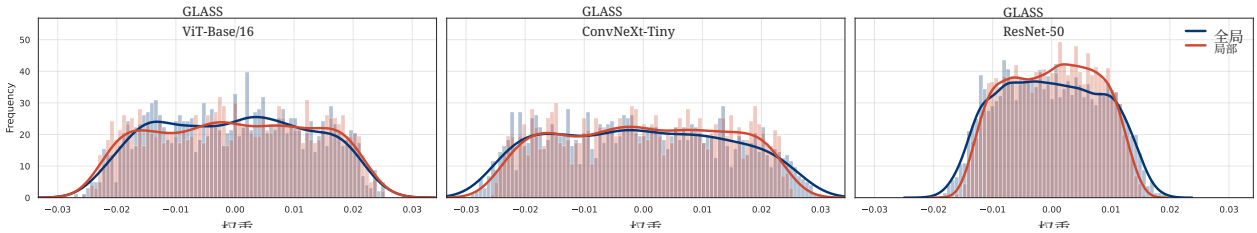

图 5：**最终基于 GLASS 的模型中分类器线性层的权重分布**。实线表示核密度估计曲线。基于 GLASS 的 ViT-Base/16 和 ConvNeXt-Tiny 模型的最终线性层各包含 3,072 个权重，而 GLASS ResNet-50 模型包含 8,192 个权重。

GLASS 架构中的分类器由一个全连接层组成，该层接收相应模型的全局与局部特征嵌入，并输出两个概率（真实和人工智能生成）。图 5展示了各模型如何在线性层中平衡全局与局部特征嵌入的贡献。全局和局部特征的权重相近，分布模式也相似；因此，它们对于分类具有同等程度的相关性和信息量。值得注意的是，GLASS ResNet-50 模型的权重分布范围似乎较窄，这可能是由于其学习率低于其他模型（超参数值见表9）。

5 结论

本文介绍的 GLASS 架构相比仅使用全局调整大小视图的标准迁移学习，准确率平均提升了 3.59%。结果表明，结合全局与局部特征能够增强模型捕捉局部细节的能力，并提升其预测性能。此外，局部采样器和注意力机制能够高效利用局部信息，而不会显著增加计算成本。总体而言，这些结果表明，GLASS 是一种很有前景的人工智能生成图像检测架构，也有望应用于涉及多尺度特征整合的其他任务。

本文存在若干局限性。首先，由于计算资源有限，数据集多样性、训练规模和超参数调优都受到显著限制。还需要进行更全面的评估，才能充分展示 GLASS 的性能。其次，所用数据集来自 2023 年，可能无法充分反映更新的生成模型的能力。最后，即使采用标准迁移学习且仅使用调整大小后的图像，也取得了较高的分类准确率。这可能表明，在当前实验设置下，这项任务相对容易。

参考文献

Akiba, T., Sano, S., Yanase, T., Ohta, T. 和 Koyama, M.（2019）。Optuna：新一代超参数优化框架。载于第 *25* 届 *ACM SIGKDD*知识发现与数据挖掘国际会议论文集。Bahdanau, D., Cho, K. 和 Bengio, Y.（2016）。通过联合学习对齐与翻译实现神经机器翻译。Chu, B., You, W., Li, M., Zheng, T., Zhao, K., Xu, X., Lu, Z., Song, J., Xu, M. 和 Zhou, L.（2025）。语义如何发挥调节作用：重新审视基于 CLIP 的生成图像检测中的图像块打乱与内部偏差。Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J. 和 Houlsby, N.（2021）。一张图像胜过 16×16 个词：用于大规模图像识别的 Transformer。He, K., Zhang, X., Ren, S. 和 Sun, J.（2015）。用于图像识别的深度残差学习。Ilse, M., Tomczak, J. M. 和 Welling, M.（2018）。基于注意力的深度多实例学习。Ju, Y., Jia, S., Ke, L., Xue, H., Nagano, K. 和 Lyu, S.（2022）。融合全局与局部特征，实现广义人工智能合成图像检测。Konstantinidou, D., Koutlis, C. 和 Papadopoulos, S.（2025）。TextureCrop：通过基于纹理的裁剪图像增强合成图像检测。Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T. 和 Xie, S.（2022）。面向 2020 年代的卷积网络。Mu, L., Xingze, Z., Bai, J., Hu, J., Zheng, W., Ye, J., Zhuang, J., Ali, M., Wang, J. 和 Hu, H.（2025）。不遗漏任何像素：一种用于稳健高分辨率人工智能生成图像检测的细节保留架构。Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, A., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J. 和 Chintala, S.（2019）。PyTorch：一种命令式、高性能深度学习库。载于神经信息处理系统进展 *32*，第 8024–8035 页。Curran Associates, Inc. Simonyan, K. 和 Zisserman, A.（2014）。用于视频动作识别的双流卷积网络。Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. 和 Salakhutdinov, R.（2014）。随机失活：防止神经网络过拟合的简单方法。机器学习研究杂志，15(56):1929–1958。Wang, S.-Y., Wang, O., Zhang, R., Owens, A. 和 Efros, A. A.（2020）。卷积神经网络生成图像竟然很容易识别……至少目前如此。Yan, S., Li, O., Cai, J., Hao, Y., Jiang, X., Hu, Y. 和 Xie, W.（2025）。对人工智能生成图像检测进行合理性检查。Zhang, T.（2023）。
ai-generated-images-vs-real-images。
h
ttps://www.kaggle.com/datasets/tristanzhang32/ai-generated-images-vs-real-images/data。Kaggle 数据集。Zhong, N., Xu, Y., Li, S., Qian, Z. 和 Zhang, X.（2024）。PatchCraft：探索纹理图像块，实现高效的人工智能生成图像检测。

A 附录

A.1 局部裁剪块的期望采样面积

给定一幅分辨率为（H， W 的图像，其中 H， $W \geq p$，且 $p \in \mathbb{N}$ 的裁剪尺寸为（p， p），我们将网格尺寸定义为

$$G = \min\left(\left\lfloor \frac{\min(H,W)}{224} \right\rfloor, 8\right) .$$

接下来，我们讨论裁剪块数量 $n \in \mathbb{N}$ 的以下情况。

情况 1： $G^2 < n$

Ea每个裁剪块的面积为 p^2，且可能存在重叠。期望总覆盖面积为

$$E[\text{covered area}] = \sum_{y=0}^{H-1}\sum_{x=0}^{W-1}\left[1-(1-p_{x,y})^n\right] \approx WH\left[1-\left(1-\frac{p^2}{WH}\right)^n\right],$$

其中 $p_{x,y}$ 表示像素 (x, y) 被包含在单个裁剪图像中的概率：

$$p_{x,y} = \frac{a_y}{H-224+1} \times \frac{b_x}{W-224+1},$$

其中 a_y 和 b_x 分别表示覆盖第 y 行和第 x列的裁剪块数量：

$$a_y = \begin{cases} y+1 & 0 \leq y \leq p-1 \\ p & p \leq y \leq H-p \\ H-y & H-p+1 \leq y \leq H-1 \end{cases} \text{ and } b_x = \begin{cases} x+1 & 0 \leq x \leq p-1 \\ p & p \leq x \leq W-p \\ W-x & W-p+1 \leq x \leq W-1 \end{cases}.$$

图像覆盖率期望值为

$$E[\text{percentage covered}] = \frac{100}{WH}\sum_{y=0}^{H-1}\sum_{x=0}^{W-1}\left[1-(1-p_{x,y})^n\right] \approx 100 \times \left[1-\left(1-\frac{p^2}{WH}\right)^n\right].$$

情况 2： $G^2 \geq n$

由于不可能发生重叠，覆盖的总面积期望值为

$$E[\text{covered area}] = np^2,$$

面积覆盖率为

$$E[\text{percentage covered}] = 100 \times \frac{np^2}{HW}.$$

A.2 超参数搜索空间

表6：**标准迁移学习模型超参数搜索空间**。方括号表示连续变量的范围，而花括号表示分类变量。所有连续值均在对数域中采样。

超参数	搜索空间
学习率	$[10^{-6}, 10^{-4}]$
权重衰减	$[10^{-7}, 10^{-3}]$
丢弃率	{0.0, 0.1, 0.2,0.3}
批量大小	{32,64, 128}

表 7：**GLASS模型超参数搜索空间**。方括号表示连续变量的范围，而花括号表示分类变量。所有连续值均在对数域中采样。

超参数	搜索空间
全局学习率	$[10^{-6}, 10^{-4}]$
局部学习率	$[10^{-6}, 10^{-4}]$
注意力和分类器学习率	$[5 \times 10^{-6}, 10^{-3}]$
全局权重衰减	$[10^{-7}, 10^{-3}]$
局部权重衰减	$[10^{-7}, 10^{-3}]$
丢弃率	{0.0, 0.1, 0.2,0.3}
批量大小	{16, 32,64}
裁剪数量	{2，4，6，8，10}

A.3 超参数调优图

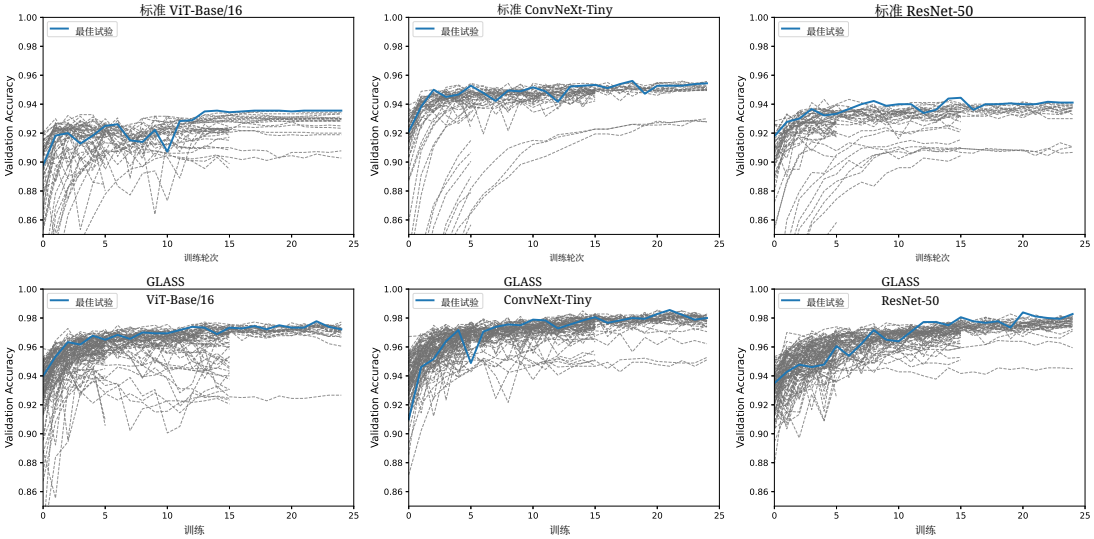

图6： **不同模型所有超参数优化试验的验证准确率曲线** .

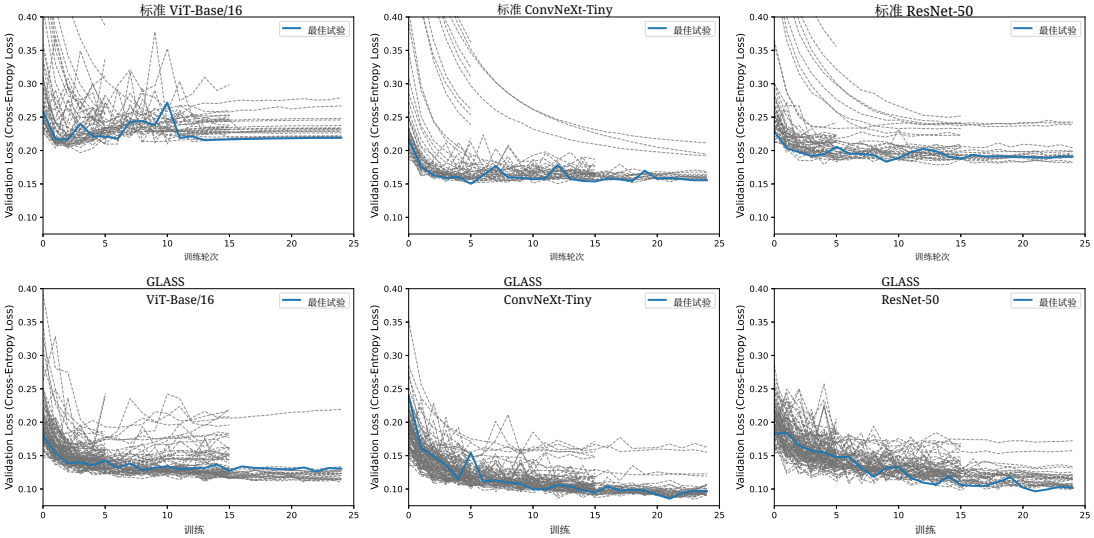

图 7：**验证 不同模态下所有超参数优化试验的交叉熵损失曲线 els.**

A.4 最佳超参数

表 8：**标准迁移学习模型的最佳超参数**。所显示的连续数值均四舍五入至三位有效数字，而最终模型使用完整精度的数值进行训练。

超参数	标准 ViT-Base/16	标准 ConvNeXt-Tiny	标准 ResNet-50
学习率	$7。48 \times 10^{-5}$	7.12×10^{-5}	$4。59 \times 10^{-5}$
权重衰减	$4。37 \times 10^{-4}$	$2。13 \times 10^{-6}$	$3。64 \times 10^{-7}$
丢弃率	0.1	0.1	0.1
批量大小	128	128	32

表9：**基于 GLASS 的模型的最佳超参数**。显示的连续值四舍五入至三位有效数字，而最终模型使用完整精度的数值进行训练。

超参数	GLASS ViT-Base/16	GLASS ConvNeXt-Tiny	GLASS ResNet-50
全局学习率	$1。58 \times 10^{-5}$	$6。55 \times 10^{-6}$	$2。16 \times 10^{-6}$
局部学习率	$4。26 \times 10^{-5}$	1.00×10^{-4}	6.85×10^{-5}
注意力和分类器学习率	6.48×10^{-5}	1.68×10^{-5}	$2。98 \times 10^{-5}$
全局权重衰减	$3。18 \times 10^{-5}$	2.85×10^{-6}	1.79×10^{-7}
局部权重衰减	6.14×10^{-6}	$3。88 \times 10^{-7}$	6.02×10^{-7}
丢弃率	0.3	0.2	0.3
批量大小	64	32	32
裁剪数量	10	4	6