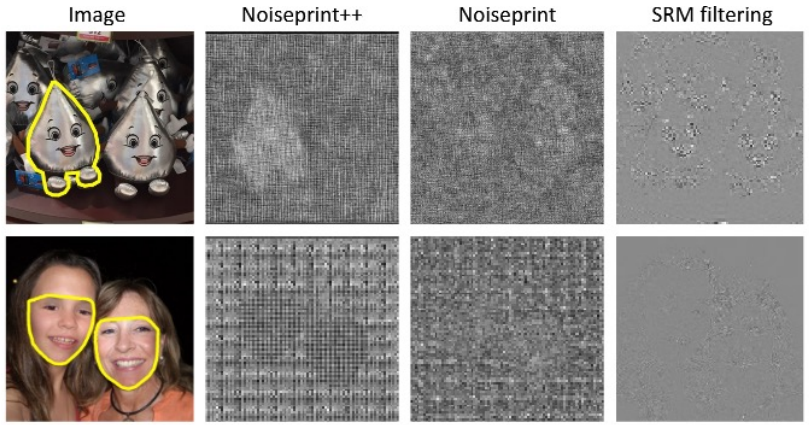
针对AI编辑图像的伪造区域定位

浙江大学网络系统安全与隐私实验室
NETWORK SYSTEM SECURITY & PRIVACY LAB

研究背景

随着扩散模型等**生成式AI**的快速发展，高质量、逼真的**AI图像编辑**变得极其容易，这加剧了虚假信息传播、身份欺诈等视觉欺骗风险。对经过AI编辑的图像进行精准识别和定位AI编辑区域成为了一个重要的问题。

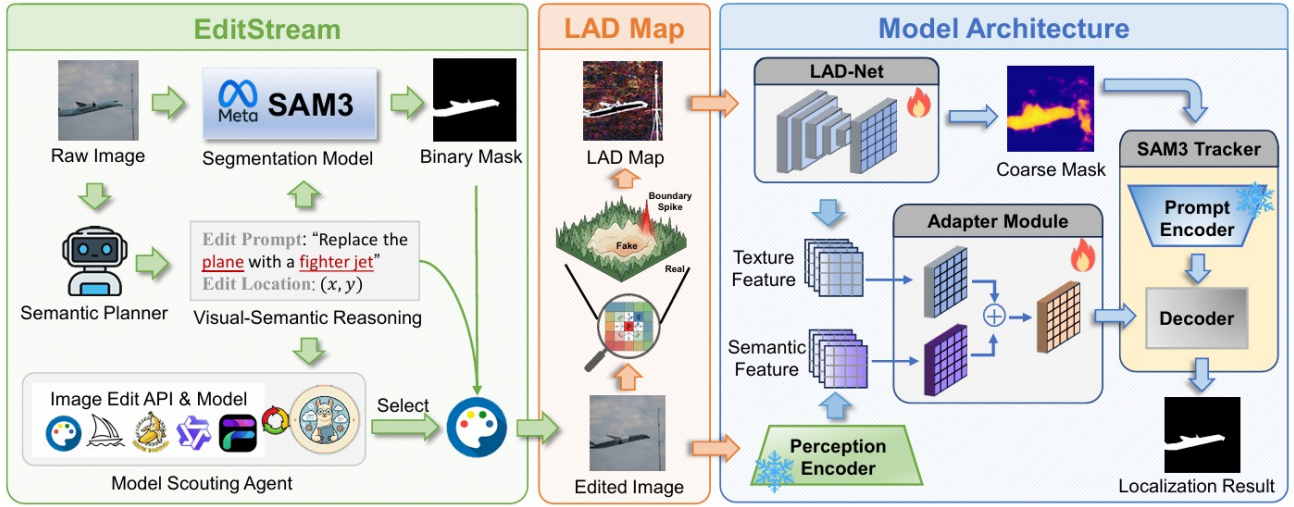
存在问题



- 传统的图像伪造定位方法主要依赖于相机传感器等物理噪声的连续性，然而**AI合成图像**中缺失这些**物理信号**，导致传统方法在面对AI生成的伪造图像时直接失效

解决方案

- **基于能量异常的定位框架**：首先利用局部邻接差异图捕捉扩散模型特有的内在**能量异常现象**，并由轻量级网络生成粗略掩码，随后结合参数高效适配器驱动的**SAM 3模型**进行语义细化，最终实现像素级的**精准伪造区域定位**。



Order within Chaos: Capturing Intrinsic Energy Anomalies for AI-Manipulated Image Forgery Localization

针对AI编辑图像的伪造区域定位

实验结果

- FLAME在多项数据集上均达到了**最佳的定位与检测水平**，准确率显著优于现有方法。
- 极强的泛化与鲁棒性：该方法不仅在应对未见过的**全新生成模型架构**时展现出卓越的泛化能力，在面对**JPEG压缩、高斯模糊和噪声**等真实场景干扰时，依然保持了高度的定位精准度

Table 1. Quantitative comparison of pixel-level localization. FLAME refers to the base model, while FLAME-F denotes the version fine-tuned via our EditStream pipeline. Underlined values indicate ID evaluation. The Average column calculates the mean performance across the last five datasets (from CoCoGLIDE to Flux Kontext) to assess generalization capabilities against OOD datasets.

Model	MagicBrush		SID		CoCoGLIDE		AutoSplice		NanoBanana		Qwen-Image		Flux Kontext		Average	
	IoU↑	F1↑	IoU↑	F1↑	IoU↑	F1↑	IoU↑	F1↑	IoU↑	F1↑	IoU↑	F1↑	IoU↑	F1↑	IoU↑	F1↑
SAFIRE	0.297	0.485	0.214	0.274	0.394	0.467	0.192	0.251	0.114	0.153	0.217	0.269	0.190	0.225	0.221	0.273
Mesorch	0.150	0.211	0.124	0.219	0.382	0.450	0.210	0.283	0.102	0.139	0.216	0.286	0.124	0.180	0.207	0.268
TruFor	0.281	0.391	0.188	0.243	0.371	0.457	0.364	0.483	0.071	0.092	0.228	0.312	0.203	0.276	0.247	0.324
AdaIFL	0.122	0.215	0.128	0.190	0.209	0.266	0.227	0.337	0.091	0.126	0.066	0.092	0.073	0.104	0.133	0.185
SIDA	0.106	0.180	0.488	0.565	0.375	0.465	0.393	0.483	0.005	0.012	0.089	0.143	0.092	0.149	0.191	0.250
FakeShield	0.091	0.126	0.117	0.137	0.138	0.150	0.238	0.296	0.086	0.095	0.098	0.110	0.096	0.108	0.131	0.152
SparseViT	0.087	0.154	0.185	0.227	0.325	0.386	0.201	0.279	0.021	0.049	0.056	0.073	0.048	0.061	0.130	0.170
FLAME (ours)	**0.538**	**0.650**	**0.580**	**0.677**	**0.469**	**0.576**	**0.501**	**0.624**	**0.216**	**0.295**	**0.321**	**0.408**	**0.285**	**0.391**	**0.358**	**0.459**
FLAME-F†(ours)	**0.507**	**0.632**	**0.569**	**0.650**	**0.481↑**	**0.602↑**	**0.498**	**0.618**	**0.391↑**	**0.454↑**	**0.482↑**	**0.603↑**	**0.446↑**	**0.548↑**	**0.460↑**	**0.565↑**

Table 2. Quantitative comparison of image-level forgery detection. The experimental setup is consistent with Table 1.

Model	MagicBrush		SID		CoCoGLIDE		AutoSplice		NanoBanana		Qwen-Image		Flux Kontext		Average	
	ACC↑	AP↑	ACC↑	AP↑	ACC↑	AP↑	ACC↑	AP↑	ACC↑	AP↑	ACC↑	AP↑	ACC↑	AP↑	ACC↑	AP↑
SAFIRE	0.525	0.640	0.596	0.570	0.481	0.481	0.462	0.623	0.570	0.592	0.454	0.644	0.616	0.492	0.517	0.566
Mesorch	0.630	0.693	0.625	0.648	0.606	0.713	0.558	0.597	0.468	0.492	0.598	0.749	0.547	0.647	0.555	0.640
TruFor	0.675	0.776	0.531	0.558	0.604	0.650	0.626	0.693	0.494	0.498	0.600	0.679	0.578	0.656	0.580	0.635
AdaIFL	0.526	0.593	0.546	0.572	0.519	0.547	0.546	0.575	0.474	0.483	0.547	0.562	0.523	0.549	0.522	0.543
SIDA	0.912	0.924	0.925	0.935	0.606	0.710	0.541	0.743	0.437	0.350	0.526	0.543	0.522	0.537	0.526	0.577
FakeShield	0.664	0.671	0.632	0.648	0.532	0.560	0.574	0.607	0.470	0.491	0.582	0.617	0.493	0.528	0.530	0.561
SparseViT	0.486	0.487	0.511	0.555	0.536	0.508	0.572	0.478	0.508	0.523	0.541	0.607	0.490	0.553	0.529	0.534
FLAME (ours)	**0.921**	**0.937**	**0.946**	**0.952**	**0.732**	**0.789**	**0.714**	**0.763**	**0.715**	**0.747**	**0.781**	**0.803**	**0.754**	**0.781**	**0.739**	**0.777**
FLAME-F†(ours)	**0.903**	**0.915**	**0.931**	**0.942**	**0.754↑**	**0.801↑**	**0.698**	**0.756**	**0.812↑**	**0.845↑**	**0.921↑**	**0.945↑**	**0.892↑**	**0.928↑**	**0.815↑**	**0.855↑**