广州大学
GuangZhou University

MUN:ImageForgery Localization Based on M3 Encoder and UN Decoder

MUN:基于M3编码器和UN解码器的图像伪造定位

Liu, Y., Chen, S., Shi, H., Zhang, X.-Y., Xiao, S., & Cai, Q. (2025). MUN: Image Forgery Localization Based on M³ Encoder and UN Decoder. *Proceedings of the AAAI Conference on Artificial Intelligence*, *39*(6), 5685-5693.

目　录

1.1 研究背景

随着数码技术的发展，图像篡改技术日益普及，伪造图像对社会安全和信任体系带来严重威胁

篡改技术普及

数码相机和图像编辑工具广泛普及,非专业用户也能制作出高度逼真的伪造图像,**技术门槛降低，篡改成本显著降低,社交媒体平台传播速度极快**

潜在风险

- 颠覆"有图有真相"的传统认知
- 传播虚假新闻，误导社会舆论
- 篡改法律证据，影响司法公正
- 操纵社会舆论，威胁国家安全

篡改类型

- **拼接**：复制粘贴不同图像内容
- **复制移动**：在同一图像内复制移动区域
- **移除**：删除特定对象或区域不同篡改类型需要特定检测方法

研究必要性

发展有效的图像篡改检测技术，以辨别原始图像内容和被篡改图像内容之间的差异，确保图像的真实性，已成为当前信息安全领域亟待解决的关键问题。

1.2 现有问题与挑战

尽管图像篡改检测领域已取得显著进展，但现有方法仍面临诸多局限性与挑战

方法局限性与通用性不足

现有方法**多为针对单一篡改类型设计**，专注于检测拼接或复制移动等**特定篡改**

特征提取不足

现有方法难以检测非语义特征（如噪声和边缘不一致），导致篡改识别不精准。

噪声特征利用不充分

尽管Noiseprint++能有效捕捉相机指纹，**如何将其部署于伪造定位并与现有技术融合，仍是尚未解决的难题。**

模型泛化能力弱

合成数据与真实图像间的分布差异，严重制约了模型的泛化能力。

这些挑战凸显了构建统一、鲁棒、高效的图像伪造定位模型的迫切需求

1.3 研究动机

针对现有方法的局限性，本研究旨在构建一个统一、鲁棒、高效的图像伪造定位模型，以应对多种篡改类型和复杂场景

统一检测多种篡改类型

新模型需能同时检测 拼接、复制移动、移除等多种篡改类型，实现统一的伪造定位。

有效融合多源特征

RGB与噪声特征的融合能更全面地表征图像中的篡改痕迹。

设计动态损失函数

传统损失函数处理不平衡数据时偏向背景像素，导致前景检测性能下降。 **IoUDCE损失函数能动态调整权重，增强对难样本的关注。**

数据增强策略

合成数据与真实图像存在分布差异，导致模型泛化能力弱。 **DNA数据增强策略通过调整RGB分布缩小差距，提升模型在复杂场景下的泛化性能。**

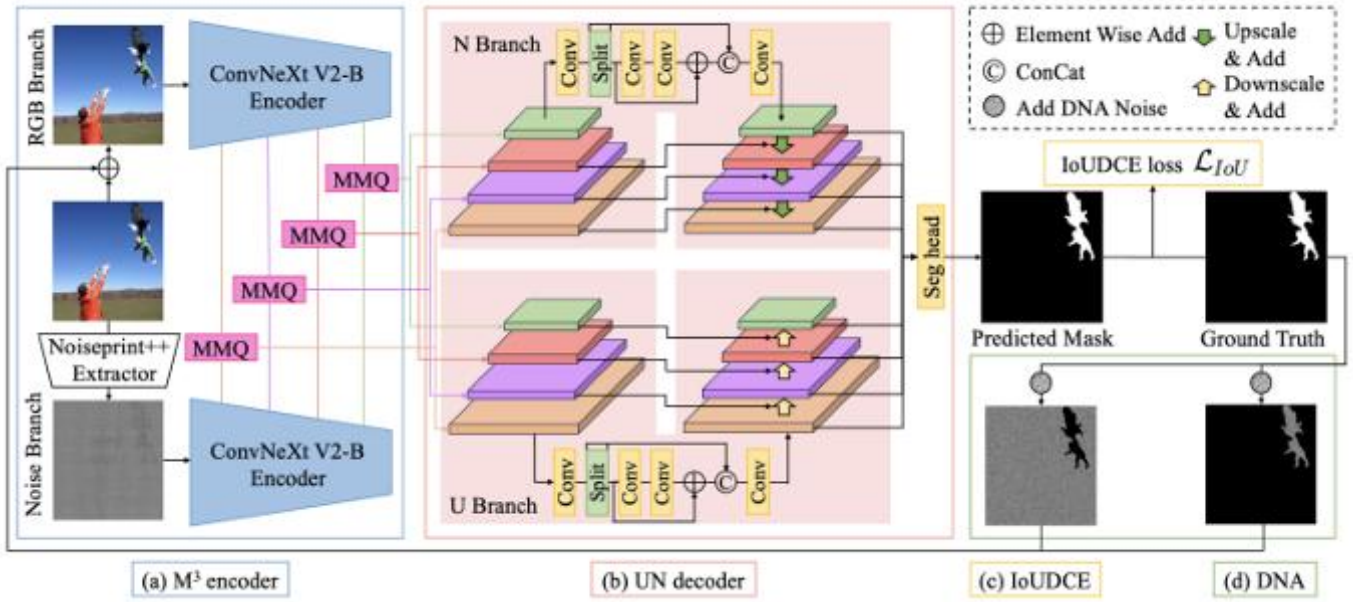
预期成果： 构建一个能在多种篡改类型、复杂场景下达到领先性能的图像伪造定位模型，有效应对现有方法的局限性。

1.4 本文贡献

- 提出**MUN网络**：基于M³编码器和UN解码器的端到端图像伪造定位框架。
- **M³编码器**：双流ConvNeXt V2结构，融合RGB与Noiseprint++特征，引入MMQ模块实现多尺度特征交互。
- **UN解码器**：双分支结构（U和N分支），分别从低层和高层特征中提取细节与语义信息。
- **IoUDCE损失**：基于IoU动态调整正负样本权重，增强对难分区域的关注。
- **DNA数据增强**：通过调整RGB分布缩小训练图像与真实图像之间的差距，提升泛化能力。
- **实验验证**：在多个公开数据集上达到SOTA性能，并在AI生成图像上表现优异。

目　录

2.1 整体框架设计



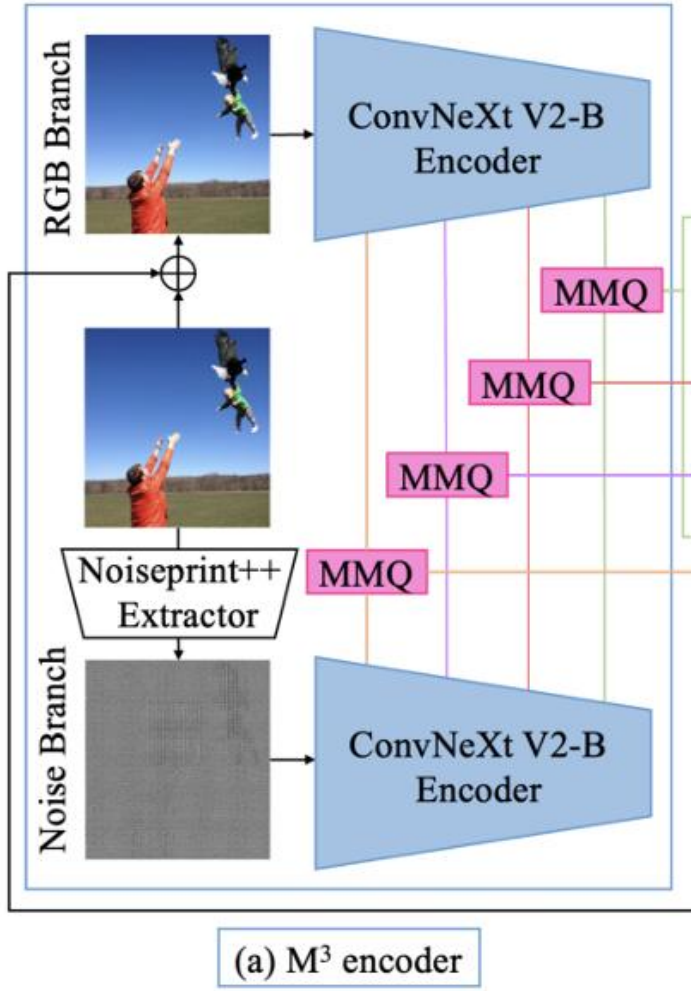
- **M³ Encoder**: 双流编码器分别提取RGB与噪声特征，并通过多尺度查询模块(MMQ)实现跨域特征融合。
- **UN Decoder**: 独特的U、N双分支结构通过自底向上与自顶向下路径协同工作，实现细节与语义特征的重建。
- **IoUDCE Loss**: 基于IoU动态调整损失权重，自适应地聚焦于难例伪造区域，解决类别不平衡问题。
- **DNA**: 通过添加基于RGB分布偏差的噪声，对齐训练集与真实数据分布，有效提升模型泛化能力。

2.2 M³编码器-双线调查

核心思想：不只看表面，更要查“指纹”

M³ Encoder采用“双线调查”策略

- ***RGB*线索流：**分析图像的**外观特征**（颜色、纹理、边缘）。就像侦探观察案发现场的整体布局和明显痕迹。
- ***Noiseprint++*线索流：**分析图像的**噪声特征**。每台相机、每次处理都会留下独特的“指纹”噪声。伪造会破坏这种一致性，留下痕迹。



(a) M³ encoder

这样做的优势

双重证据，相互印证。

2.3 M³编码器组成部件

（1）选择强大的“调查员”——骨干网络ConvNeXt V2

- **性能强劲：** 在多项测试中，其精度超过传统的ResNet和Transformer风格的Swin网络。
- **效率高超：** 在达到更高精度的同时，计算成本更低。
- **更关注局部：** 伪造痕迹往往是局部的不一致，ConvNeXt的卷积结构天生擅长捕捉

（2）Noiseprint++的部署（提取有效的“噪声指纹”）

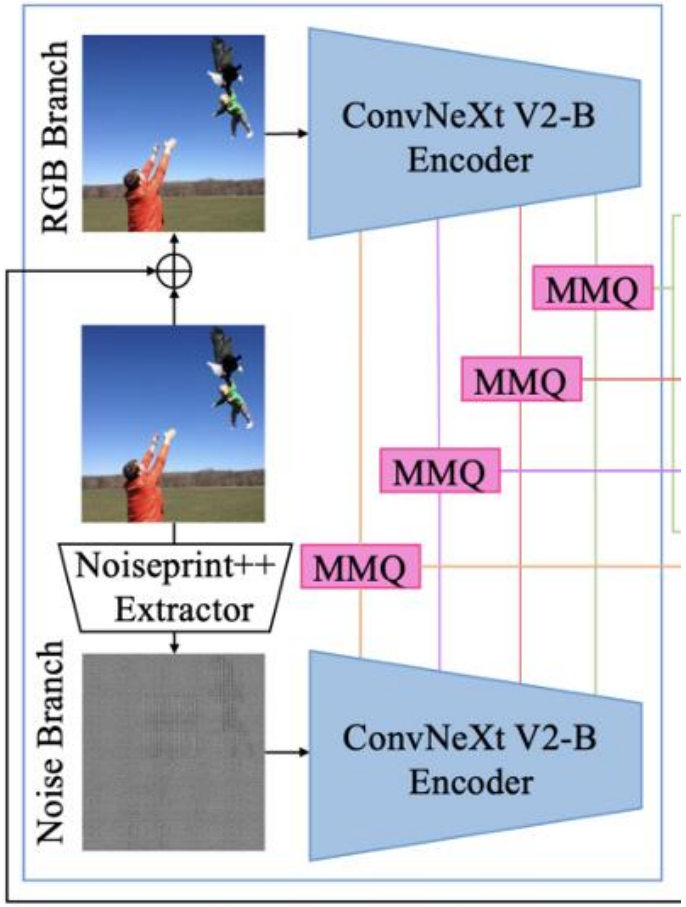
问题： 图像输入网络前需要调整大小（Resize），但这个操作本身会干扰噪声特征。

探索部署策略

① 只调整RGB图大小（忽略噪声）

② 先调整图大小，再提取Noiseprint++（噪声被污染）

③ 先提取原图的Noiseprint++，再调整图大小（胜出！）



(a) M³ encoder

2.4 M³编码器的核心-MMQ模块

问题： 简单地把两条线索拼接在一起太粗暴，无法建立深层次关联

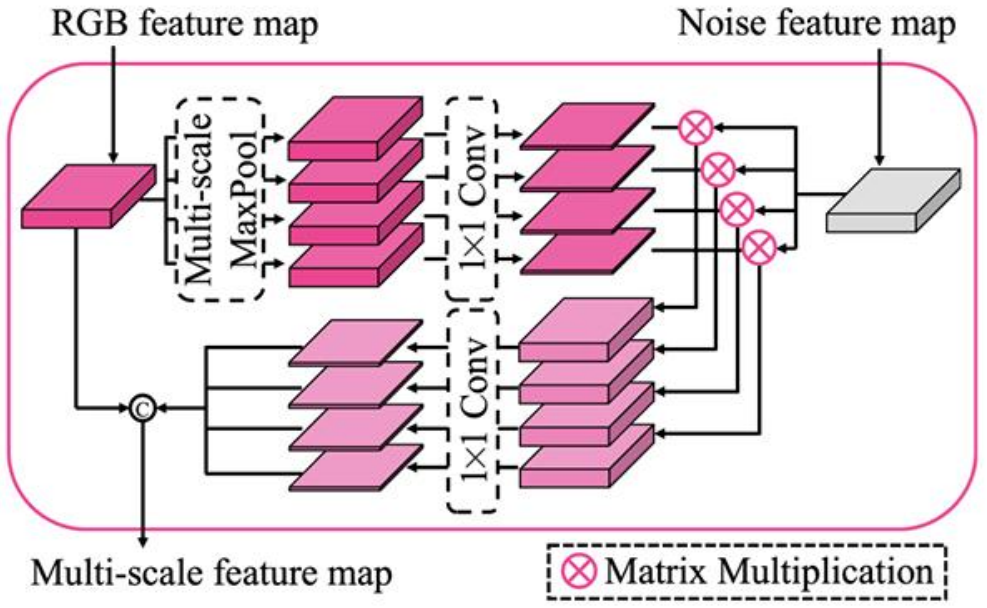
核心创新：让两条线索对话——MMQ模块

MMQ模块

① **提问 (*Query*)**：从*RGB*特征中，用不同大小的窗口（*1x1*，*5x5*，*11x11*，*15x15*）进行**最大池化**，**提取不同尺度下“最显著”的局部特征作为“问题”。**

② **检索 (*Retrieval*)**：将这些“问题”与*Noiseprint++***特征图**进行矩阵乘法操作，**相当于在噪声域中寻找与这些显著*RGB*区域最相关的“答案”。**

③ **融合 (*Fusion*)**：将不同尺度下找到的“答案”与原始*RGB*特征**拼接**起来，形成丰富的多线索、多尺度特征。



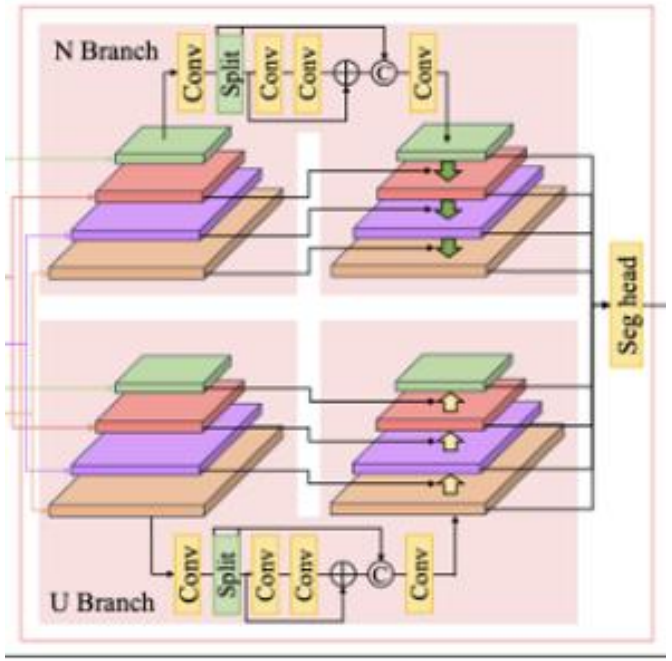
2.5 UN解码器的设计理念

回顾：编码器提供的4份（特征图）是不同抽象层次的

- **低层特征 (如 Q(0))**: 细节丰富（如边缘、纹理），但缺乏全局语义信息。
- **高层特征 (如 Q(3))**: 语义信息强（知道哪里“有问题”），但缺乏细节，非常模糊。

***UN*解码器的策略：** 开辟两条并行的“工作线”（分支），同时从高层和低层特征开始处理，双向奔赴，最终融合。

- ***U*分支 (*Up Branch*): 自下而上。**从**低层**特征出发，主要负责**恢复细节、精修边界。**
- ***N*分支 (*Down Branch*): 自上而下。**从**高层**特征出发，主要负责**传递语义信息、锁定可疑区域。**



(b) UN decoder

2.6.1 U分支：自下而上，精修细节

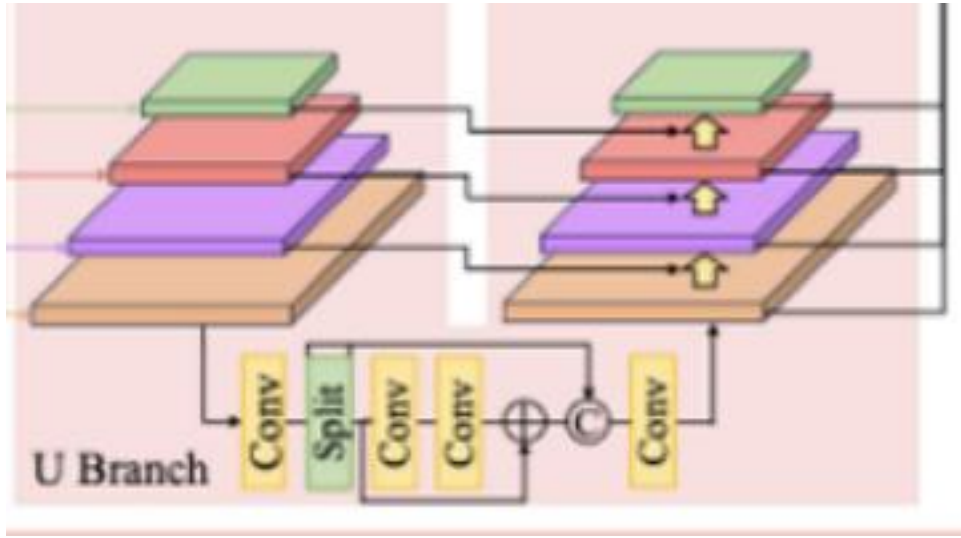
U分支：从细节出发，逐步融合全局信息

工作流程 (Bottom-Up)

① **起点：** 对最底层的细节特征 $Q(0)$ 进行一系列轻量级卷积操作，提取更精细的**细节特征** $BU(0)$。

② **向上融合：** 将 $BU(0)$ **上采样**后，与上一层的特征 $Q(1)$ **相加融合**，得到既包含本层细节又包含上一层语义的 $BU(1)$。

③ **重复过程：** 此过程重复，直到最顶层。最终，U分支输出了4个融合了自下而上细节信息的特征图 BU。



2.6.2 N分支：自上而下，传递语义

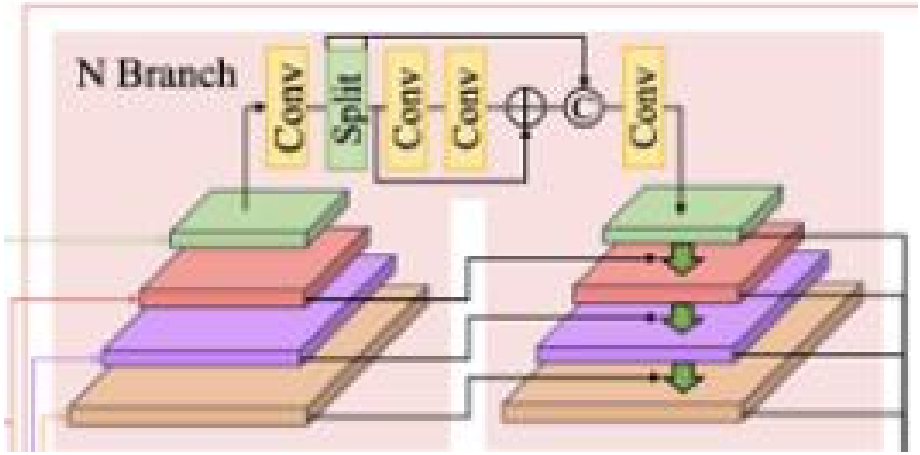
N分支：从全局出发，逐步指导细节恢复

工作流程 (Top-Down)

① **起点：** 对最高层的语义特征 *Q(3)* 进行一系列轻量级卷积操作，提炼更纯粹的**语义特征** *TD(3)*。

② **向下融合：** 将 *TD(3)* **上采样**后，与下一层的特征 *Q(2)* **相加融合**，将高层语义信息“注入”到更底层的特征中，得到 *TD(2)*。

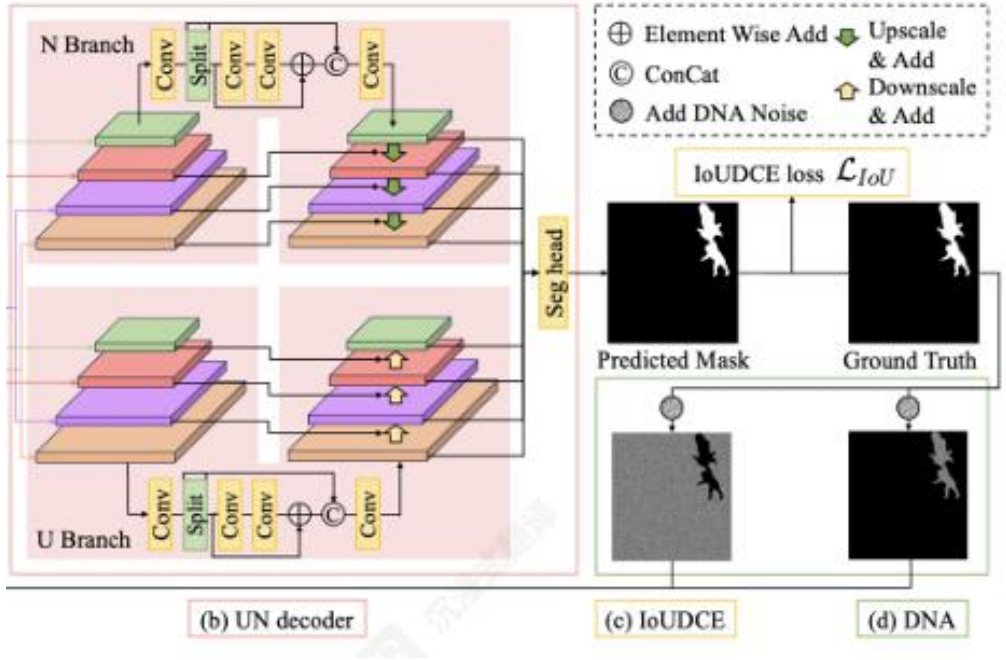
③ **重复过程：** 此过程重复，直到最底层。最终，*N*分支输出了4个被高层语义信息增强过的特征图 *TD*。



2.6.3最终融合：绘制“造假地图”

双线合璧，生成高精度预测结果

- **输入：** U分支输出的 *BU* 特征图集合（富含细节）和 N 分支输出的 *TD* 特征图集合（富含语义）。
- **融合过程：**
  - **层级对接：** 将同一层次（如都是第*O*层）的 *BU(i)* 和 *TD(i)* 特征图**拼接（*Concat*）**在一起。
  - **卷积调整：** 通过卷积层调整融合后的特征，得到该层最终的融合特征 *UN(i)*。
  - **统一尺寸：** 将所有层次的 *UN(i)* 特征图上采样到相同的尺寸（与最底层相同）。
- **输出结果：** 将所有统一尺寸的融合特征再次拼接，最后通过一个简单的卷积层，像“盖章”一样，生成最终的**二值预测掩膜**——清晰的“造假地图”。图中白色区域即为模型预测的伪造区域。



2.5.2 IoU重构动态交叉熵损失 (IoUDCE Loss)

动机: 篡改区域与真实背景存在极端的像素数量不平衡，需动态调整损失权重以聚焦于难例（篡改区域）

方法

① **计算*Batch IoU* ($\boldsymbol{\beta}_{\mathrm{IoU}}$)**: 在一个训练批次(*Batch*)内，计算预测掩膜与真实掩膜的平均交并比(*IoU*)。

② **动态权重重构**: 基于$\boldsymbol{\beta}_{\mathrm{IoU}}$动态重构正样本（篡改像素）的损失权重。权重因子为 $e^{\lambda(1-\boldsymbol{\beta}_{\mathrm{IoU}})}$。

- 当 $\boldsymbol{\beta}_{\mathrm{IoU}}$低（预测差），($1$ - $\boldsymbol{\beta}_{\mathrm{IoU}}$) 值大，权重因子显著增大，**强化对篡改区域的训练监督。**
- 当 $\boldsymbol{\beta}_{\mathrm{IoU}}$高（预测好），权重因子趋近于$1$，损失函数退化为平衡的交叉熵形式。

$$\beta_{IoU} = \sum_{i}^{B} \frac{\mathbf{GT}_b(i) \cap \mathbf{M}_b(i)}{\mathbf{GT}_b(i) \cup \mathbf{M}_b(i)} \tag{8}$$

$$\begin{aligned}\mathcal{L}_{IoU} = -\frac{1}{B}\frac{1}{HW}\sum_{i=1}^{B}\sum_{j=1}^{HW}(&\alpha_0(1-y(i,j)) \\ &\times \log(1-p(i,j)) + \alpha_1 \\ &\times e^{\lambda(1-\beta_{IoU})} y(i,j)\log p(i,j))\end{aligned} \tag{9}$$

其中 B 表示批次大小，$\cap$ 和 $\cup$ 分别表示交集和并集，$y(i,j)$ 表示批次中第 i 张图像的第 j 个真实标签，$p(i,j)$ 表示批次中第 i 张图像的第 j 个伪造预测分数。α_0、α_1 和 λ 是超参数。

优势

该损失函数能够**自适应地**根据模型当前性能调整优化重点，有效促进模型对难例的学习

2.5.2 DNA数据增强

核心思想: 通过向训练图像注入特定设计的噪声，将其*RGB*分布向更具代表性的**通用图像分布**（如*ImageNet*）进行对齐，从而在训练阶段模拟域间差异。

① 计算偏差

- 计算训练集与*ImageNet*的***RGB*均值偏差向量**$\mathrm{d} = \mathrm{m}^{\mathrm{u}} - \mathrm{m}^{\mathrm{t}}$。

② 生成加权噪声

- 生成均匀噪声N_{p}，N_{f}。
- 按通道用偏差 d 进行加权：$\hat{\mathrm{N}} = \mathrm{d} \circ \mathrm{N}$。

③ 区域自适应注入

- 根据真实掩膜 *GT*，将噪声$\widehat{\mathrm{N}_{\mathrm{p}}}$与$\widehat{\mathrm{N}_{\mathrm{f}}}$分别添加到**真实区域**与**篡改区域**。

$$\mathbf{I}_{DNA} = \bar{\mathbf{I}}_{RGB} + (\hat{\mathbf{N}}_p \circ (\hat{\mathbf{GT}}_b = 0)) + (\hat{\mathbf{N}}_f \circ (\hat{\mathbf{GT}}_b = 1))$$

目　录

3.1 实验设计

- **实验目标：**验证MUN网络在图像伪造定位任务中的**有效性、泛化性与鲁棒性。**
- **实验内容**
  - **消融实验：**骨干网络评估，验证M3编码器、UN解码器、IoUDCE损失、DNA增强等模块的有效性。
  - **泛化性能测试：**在多个公开数据集（NIST16、CASIA、IMD2020、CocoGlide、Wild）上测试模型泛化能力。
  - **对比实验**：与RGB-N、ManTra-Net、SPAN、MVSS-Net、PSCC-Net、ObjectFormer、TANet、TBFormer、HiFi、TruFor、CSR-Net、NRL-Net、MGQFormer等SOTA方法对比。
  - **鲁棒性测试：**评估模型对图像缩放、模糊、JPEG压缩等后处理操作的鲁棒性。
- **数据与评估**
  - **数据集**
    - 训练集：Synthesized Dataset（基于CASIA v2.0 和 ADE20k，包含拼接、复制-移动、移除篡改）
    - 测试集：NIST16、CASIA v1.0、IMD2020、CocoGlide（AI生成）、Wild
  - **评估指标**：F1-Score、IoU、Accuracy、AUC（阈值固定为0.5）

3.2 骨干网络评估

输入: 均为 512 × 512

对比: Swin-B, ResNet-152, ConvNeXt V2-B

Backbone	F1	FLOPS (G)	Params (M)
ResNet-152	0.9326	123.988	91.233
Swin-B	0.9450	109.568	90.324
ConvNeXt V2-B	0.9519	94.204	91.139

Table 1: Backbone evaluation on Synthesized Dataset

由于ConvNeXtV2- B**在计算复杂度更低且参数相似的情况下**实现了比Swin- B和ResNet- 152更高的F1分数， 我们选择ConvNeXtV2- B作为骨干网络。

3.3 MUN组件有效性

- “RGB”表示单个ConvNeXtV2分支从RGB图像中提取线索的版本；
- “RGB+IoU”意味着我们使用IoUDCE损失而不是交叉熵损失
- “RGB+IoU+UN”表示构建UN分支的版本；
- “RGB+IoU+UN+NPP”在噪声分支中直接连接两个分支的中间特征图。

Variants	F1	IoU	Acc	AUC
RGB	0.9519	0.9156	0.9943	0.9985
RGB+IoU	0.9523	0.9161	0.9943	0.9985
RGB+IoU+UN	0.9525	0.9166	0.9943	0.9986
RGB+IoU+UN+NPP	**0.9533**	**0.9170**	**0.9945**	**0.9989**

Table 2: MUN ablation study on Synthesized Dataset

结论：MUN的各个组成部分均对网络性能有积极作用

3.4 MMQ参数设置

Kernel size	F1	IoU	Acc	AUC
None	0.9533	0.9170	0.9945	0.9989
1	0.9540	0.9184	0.9947	0.9989
1, 5	0.9540	0.9183	0.9939	0.9989
1, 5, 11	0.9542	0.9186	**0.9947**	0.9989
1, 5, 11, 15	**0.9543**	**0.9186**	0.9946	**0.9989**

Table 4: MMQ parameter setting on Synthesized Dataset

当MMQ具有单个1×1 最大池化流时，F1分数增加了0.07%。使用更多不同核大小的最大池化流，性能可以得到进一步改进。**MMQ可以指导MUN学习多尺度RGB特征和Noiseprint++特征之间的相关性。通过最大池化操作，一些不太重要的信息可以被忽略，而MUN更加关注不同尺度下局部特征的重要信息。**

3.5 验证鲁棒性

我们考虑了三种常见的后处理方法用于鲁棒性评估：
(1)使用两个较小的比例调整伪造图像的大小，
(2)使用两个不同尺寸的卷积核对伪造图像进行平滑处理
(3)使用两个不同的质量因子压缩伪造图像。

Distortions	MVSS-Net	MUN
no distortions	0.814	**0.885**
Resize(0.78×)	0.799 (-0.015)	**0.874** (-0.011)
Resize(0.25×)	0.700 (-0.114)	**0.835** (-0.050)
GaussianBlur(k=3)	0.796 (-0.018)	**0.865** (-0.020)
GaussianBlur(k=15)	0.761 (-0.053)	**0.813** (-0.072)
JPEGCompress(q=100)	0.811 (-0.003)	**0.888** (+0.003)
JPEGCompress(q=50)	0.786 (-0.028)	**0.818** (-0.067)

Table 5: Robustness evaluation on IMD2020

可以看出MUN始终取得了更好的性能，这表明MUN具有良好的鲁棒性。

3.6 与SOTA方法对比

Methods	NIST16	CASIA v1.0	IMD2020	CocoGlide	Wild
RGB-N$^{CVPR'18}$	0.764	0.795	-	-	-
ManTra-Net$^{CVPR'19}$	0.795	0.817	0.748	0.778	0.677
SPAN$^{ECCV'20}$	0.840	0.797	0.750	0.475	-
MVSS-Net$^{ICCV'21}$	-	0.815	0.814	0.654	0.768
PSCC-Net$^{TCSVT'22}$	0.855	0.829	0.806	0.777	0.745
ObjectFormer$^{CVPR'22}$	0.872	0.843	0.821	-	-
TANet$^{TCSVT'23}$	<u>0.898</u>	0.853	0.849	-	<u>0.832</u>
TBFormer$^{SPL'23}$	0.847	0.955	0.863	0.747	0.783
HiFi$^{CVPR'23}$	0.869	0.866	0.834	-	-
TruFor$^{CVPR'23}$	0.839	0.833	0.818	0.752	-
CSR-Net$^{AAAI'24}$	0.883	0.881	0.854	-	-
NRL-Net$^{AAAI'24}$	**0.900**	0.872	0.852	-	-
MGQFormer$^{AAAI'24}$	0.862	0.886	0.883	-	-
MUN	0.857	**0.967**	<u>0.885</u>	<u>0.811</u>	0.805
MUN*	0.861	<u>0.962</u>	**0.897**	**0.815**	**0.843**

The bold entities denote the best results per column and the underlined ones denote the second best results. * denotes that the DNA data augmentation method is performed. AUC scores are reported.

Table 6: Comparison against the State-Of-The-Art methods

MUN在大多数数据集上表现更好。

目　录

4.1 总结

- **方法创新**：提出MUN网络，结合M3编码器（RGB+Noiseprint++双流）与UN解码器（双向特征融合），并设计MMQ模块增强跨模态特征关联。
- **关键模块**：
  - M3编码器：利用ConvNeXt V2提取RGB与噪声特征，通过MMQ实现多尺度池化查询融合。
  - UN解码器：并行U（低阶细节）与N（高阶语义）分支，双向融合生成预测掩码。
- **损失与增强**：提出IoUDCE损失，动态调整伪造区域权重；设计DNA增强，缩小训练与真实图像RGB分布差异。
- **实验验证**：
  - 在多个数据集（NIST16、CASIA、IMD2020等）上F1/IoU优于SOTA，支持AI生成图像检测。
  - 消融实验验证各模块有效，鲁棒性测试抗缩放/模糊/压缩攻击性能领先。

感谢聆听

欢迎老师、同学们批评指正