# 全防御框架：多层次深度伪造检测与溯源

石慧[1]，王桂宾[1]，李彦妮[2]，戚茹佳[1]
[1]辽宁师范大学计算机与人工智能学院，中国大连市，116021
[2]辽宁对外经贸学院管理学院，中国大连市，116029

**摘要：**深度伪造已对政治、新闻、娱乐等多个领域构成严重威胁。尽管大量基于被动检测或主动防御的方法已被提出，但很少有方法能够同时实现被动检测和主动防御。为解决这一问题，我们提出一种基于交叉域特征融合和可分离水印的全防御框架，同时实现被动检测和主动防御。主动防御模块由一个编码器和两个可分离解码器组成，其中编码器将水印嵌入到受保护的人脸图像中，两个解码器分别提取具有不同鲁棒性的水印。鲁棒水印能够可靠地追踪可信的人脸，而半鲁棒水印则对恶意攻击（深度伪造攻击或水印移除攻击）敏感，这些恶意攻击会导致水印消失。当水印消失时，被动检测模块则融合空间域和频率域特征，进一步区分到底是经过了深度伪造攻击还是水印移除攻击。所提出的交叉域特征融合策略首先用频率域特征的“主要”通道替换空间域特征的“次要”通道，再用空间域特征的“主要”通道替换频率域特征的“次要”通道。大量实验表明，所提出的方法不仅提供主动防御机制（即溯源和版权保护），还在无水印的情况下实现被动检测，进一步区分深度伪造攻击和水印移除攻击，从而提供全面的防御框架。

**关键词：**深度伪造检测；主动防御；溯源；交叉域特征融合；水印移除攻击
https://doi.org/10.1631/FITEE.2401012