Xiaochen Yuan received the Ph.D. degree in Software Engineering from the University of Macau in 2013. From 2014 to 2015, she was a postdoctoral fellow at the Department of Computer and Information Science of the University of Macau. From 2016 to 2021, she was an Assistant Professor and an Associate Professor at the Faculty of Information Technology of the Macau University of Science and Technology. Since 2021, she joined the Faculty of Applied Sciences of the Macao Polytechnic University, where she is currently an Associate Professor. Her research interests include Multimedia Forensics and Security, AI Model Security, Quantum Watermarking, Remote Image Processing, and Deep Learning Techniques and Applications. She has published more than 80 SCIE-indexed scientific articles in refereed journals, such as IEEE TIFS, IEEE TII, IEEE TMI, IEEE TIM, IEEE TETC, IEEE JSTARS, etc. and she has been served as reviewers for top-tier journals and conferences in related areas, such as IEEE TIFS, IEEE TMM, IEEE TCSVT, IEEE TIM, IEEE TDSC, CVPR, ICME, etc. She also serves as the program committee member, session chair, and regional chair for ICSPS, ICSIP, ICSTE, etc. She is also a member of CCF, a member of CSIG, a member of IET, and a senior member of IEEE.

袁小晨，现任澳门理工大学应用科学学院副教授，博士生导师，IEEE 高级会员。于 2013 年获得澳门大学软件工程博士学位，2013 年 11 月至 2015 年 10 月于澳门大学科技学院图像处理与模式识别实验室担任博士后研究员，2016 年 1 月至 2021 年 7 月于澳门科技大学资讯科技学院担任助理教授及副教授，从事教学科研工作。于 2021 年 8 月加入澳门理工大学应用科学学院担任副教授。主要研究领域包括多媒体安全和取证、数字水印技术、人工智能模型安全、量子水印、遥感图像处理以及人工智能和深度学习技术及应用等。在国际权威期刊及学术会议上发表学术论文逾

百篇，其中 SCI/SCIE 收录 90 余篇。申请/授权中国发明专利 15 项。主持/完成研究项目 7 项，包括国家自然科学基金青年项目、澳门科学技术发展基金科研资助项目等。担任多个著名学术期刊的审稿人，以及多个国际会议的程序委员会委员，分会场主席和地区主席。

AI-Driven Protection for Content and Models

Abstract

Addressing the urgent challenges of digital content authenticity and generative AI model security in the current AI era, we are dedicated to research topics of multimedia security and forensics, spanning from content to model. At the digital content level, leveraging AI, we develop techniques for detecting both superficial manipulations and sophisticated deepfakes in images, precisely pinpointing manipulated regions. This provides reliable technological means to ensure the authenticity and trustworthiness of digital content. At the model protection level, we integrate white-box watermarking mechanisms to safeguard against parameter tampering risks. Concurrently, we employ black-box watermarking technology to enable copyright tracing and infringement verification without disclosing the model's internal details. Through the synergistic integration of these two critical domains, we target at significantly enhancing the security assurance for both AI-generated content and AI models themselves, and establishing a robust security foundation and bolstering trustworthiness within the AI technology ecosystem.