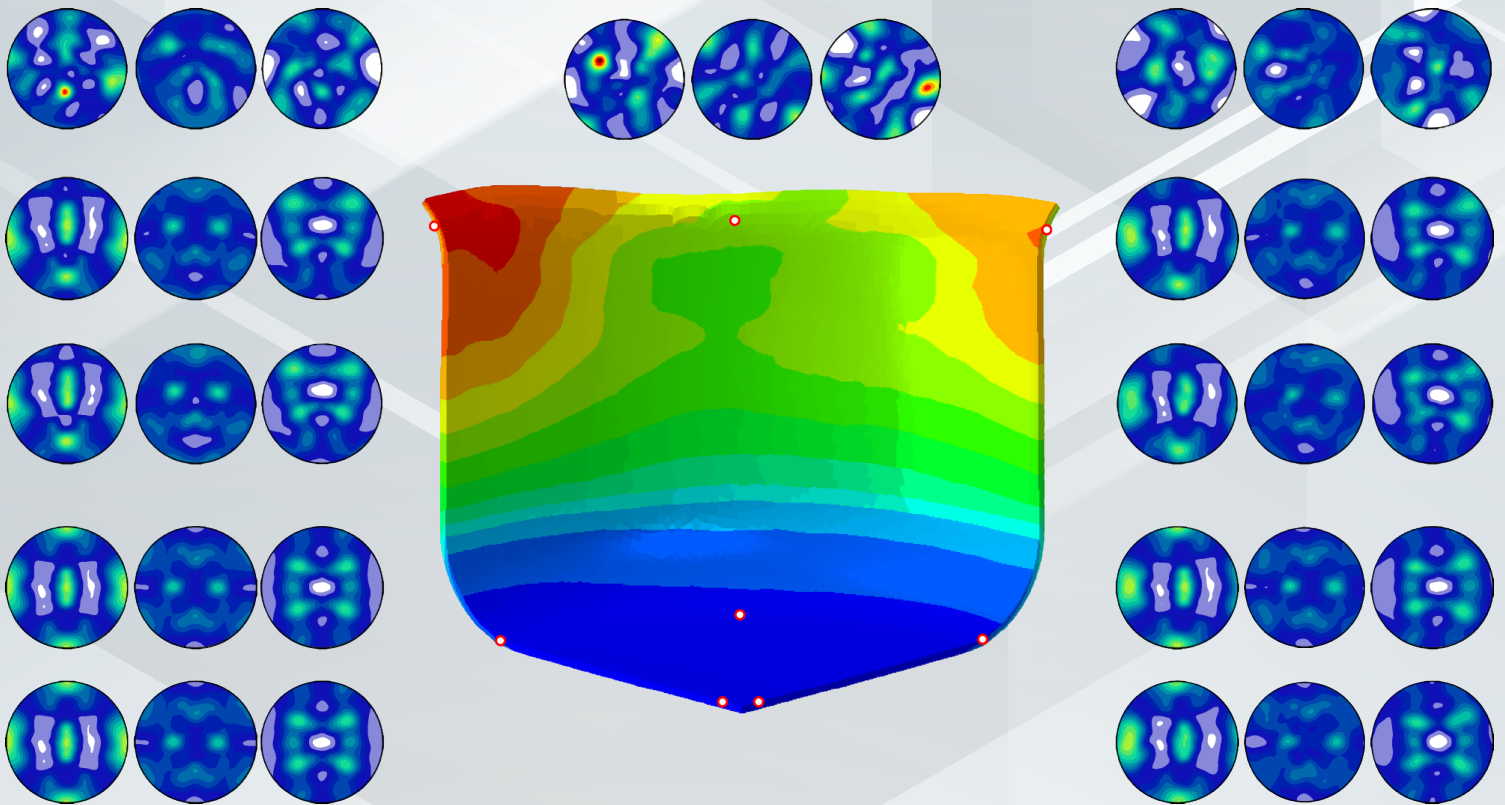


Instruments Today 2026 Q3

科儀新知



先進人工智慧技術與應用

Advanced AI Technologies and Applications

【人物專訪】

旺宏電子前瞻技術實驗室奈米技術

謝光宇 資深處長

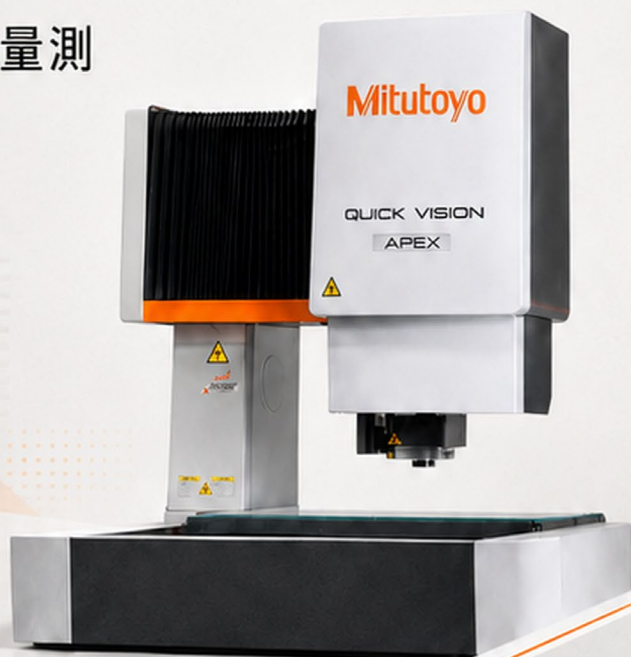
248

ISSN 1019-5440

Mitutoyo

限時回饋促銷

影像量測



形狀量測



三次元量測



詳情請洽台灣三豐儀器各地辦事處

台北營業部 (02)5573-5900

台中辦事處 (04)2338-6822

高雄辦事處 (07)334-6168

台灣三豐儀器官方網站



GMM

均華精密工業股份有限公司

Gallant Micro. Machining Co., Ltd.

(股票代號：6640)

2010

總部成立於土城工業區內

均華以尖端技術引領半導體AI晶片製造領域，特別在先進的CoWoS封裝製程設備上展現出卓越的技術能力。

2018

正式掛牌上櫃(6640)

主力產品設備包括頂尖的精密取放技術之高效能的晶粒挑揀機與黏晶機。藉由先進的精密加工技術實現高加工精度和速度的沖切成型機與自動封膠機，同時在光電整合技術方面，我們的雷射刻印機與光學檢測機為產業界指標性規格。

主要商品/服務項目

半導體先進封裝精密取放設備設計、製造、銷售

自動雷射刻印機、自動沖切成型機設計、製造、銷售

精密模具及機械設備之設計、製造、銷售

半導體先進封裝、異質整合

歡迎各界菁英加入 共創產業未來

★完善的獎勵

年終獎金、員工分紅(視公司年度營運狀況、個人績效表現分發) / 三節禮金 / 生日禮金 / 庫藏股

★完善教育訓練

新進人員基礎訓練(師徒制) / 強化研發技術能量(內部教育訓練課程) / 個人潛能提升課程

★完整的保險福利

勞工保險 / 勞工退休金(新制) / 全民健康保險 / 公司團體保險(壽險、意外傷害險、意外醫療險、住院醫療險、癌症醫療險) / 員工及眷屬保險(提供優惠方案給予眷屬納保)

★休假

全薪病假3天 / 生理假 / 產檢假 / 陪產假

★職工福利

公司及福委會提供結婚禮金、生育禮金、喪葬補助金 / 員工家庭日 / 部門聚餐、歲末尾牙餐會 / 員工健康檢查 / 哺乳室 / 知識分享獎金 / 員工子女獎助學金 / 提供書刊及雜誌借閱 / 國內外旅遊 / 彈性上下班制度 / 社團活動

GMM

236 新北市土城區民生街2-1號
TEL：02-2268-2216 FAX：02-2268-0665

302 新竹縣竹北市台元一街8號7樓之9
TEL：03-5601222 FAX：03-5601292



高科技領導廠商

CONTROL

東捷科技



專注研發核心技術

自動化整廠規劃與製程設備

雷射應用設備

光學檢測設備

物流規劃與設備

真空製程設備

巨量轉移/巨量修補設備

以研發為主的設備製造商
專注於新型雷射應用與雷射微加工技術

擅長LCD製程設備整合整廠自動化

近年將技術應用擴展至 (micro) LED及半導體產業

為國內高科技領導廠商!

完善服務網絡，服務據點在地化
維修保養、定期設備調校、客戶問題即時對應！

WWW.CONTREL.COM.TW

☎ 06-5051188(台南總公司) 📍 服務據點:新竹.台中.台南.高雄

關注東捷



TILTdot

無線傾斜計

TILTdot 是一款具備 LTE-M 與 NB-IoT 無線傳輸功能的電子式傾斜儀，提供雙軸角度與溫度量測。採用堅固防水外殼與低功耗設計，搭配 2 顆 18650 鋰電池即可進行長時間量測，並將資料上傳至 dot 雲端監測平台。內建 32 GB Micro SD 卡，可備份超過一百萬筆資料，確保量測資料完整可靠。

- 大樓結構物傾斜監測
- 橋樑墩柱傾斜監測
- 隧道結構傾斜監測
- 邊坡擋土牆傾斜監測
- 地表滑動監測

應用環境：



都市土木新建工程



河川水利工程



橋梁工程



邊坡安全



pALERT F330

一體化力平衡式地震儀

本設備搭載高精度三軸力平衡加速度計，結合地震預警技術，可即時偵測地震波並快速觸發預警機制，有效降低災害衝擊與風險。儀器提供 Pd、PGA 及 STA/LTA 三種觸發邏輯，並內建濾波機制，可有效降低非自然地震噪音干擾，完整記錄高品質地震波形，供進一步分析與事件判釋使用。

- 地震預警系統(EEWS)
- 工業風險管理增強
- 結構健康監測(SHM)
- 生產線地震安全停機機制

應用環境：



住宅建築



工業基礎設施



製造廠

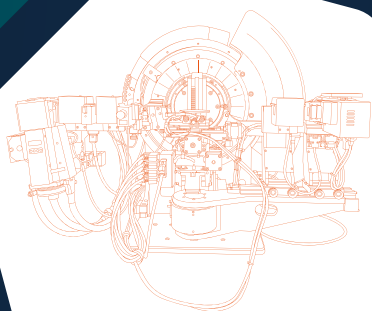


公共設施

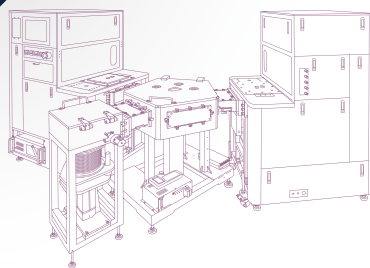




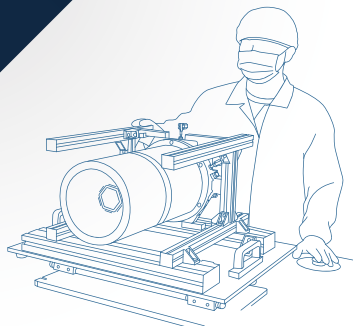
Frontier Technology Dream Incubator



In-plane X-ray Diffraction Analysis
垂直面 X-ray 繞射分析儀



Cluster Atomic Layer Deposition System
叢集式 ALD 製程設備



Meter-scale Optical System
航太級大口徑光學系統



Intelligent Instrument Design & System Integration

智慧儀控設計與整合



Total Solution Service for Optics

光學系統設計、製造與檢測



Advanced Coating Equipment Development

先進鍍膜製程設備開發



Biomedical Optical System

生醫光電系統



<https://www.ncir.niar.org.tw/>



NCIR 國家實驗研究院

國家儀器科技研究中心

National Center for Instrumentation Research

新代科技



TRUSTED • TECHNICAL • SERVICE

智慧領航：AI 驅動新世代製造方案

AI-Driven Solutions for Next-Gen Manufacturing



機器人

SCARA
桁架機器人
關節機器人
協作機器人

工具機

車床
銑床
五軸機
玻璃高光機

機聯網

IOT
MES
QMS
ESG

伺服

主軸馬達
驅動器
直驅馬達
編碼器

自動化

雷射切割
雷射焊接
雷射打標
磨床

AI



新代科技股份有限公司
SYNTEC Technology Co., Ltd.
台灣新竹市東區研發二路25號

☎ 886-3-6663553
☎ 886-3-6663505
✉ sales@syntecclub.com.tw

新代科技(蘇州)有限公司
Suzhou SYNTEC Equipment Co., Ltd.
江蘇省蘇州工業園區春輝路9號

☎ 86-512-69008860
☎ 86-512-65246029
✉ service@syntecclub.com.cn

新代科技



聯達智能



YouTube



LinkedIn



*SYNTEC 保有隨時修改型錄資訊的權利，因印刷錯誤或任何錯誤修訂恕不另行通知。

©2026.05 (PRINTED IN TAIWAN)



中華民國儀器商業同業公會全國聯合會
Taiwan Instruments United Association

全國儀器產業 領航者 攜手共進新紀元

「中華民國儀器商業同業公會全國聯合會」自 107 年成立以來，已成為台灣儀器產業最具代表性的樞紐。我們凝聚了北、中、南共 6 大地方儀器公會的力量，旗下匯流逾 1,200 家企業會員，深耕全台各個角落。

本會致力於搭建同業溝通橋樑，推動國內外貿易拓展，共創產業最大福祉。會員企業專精於理化分析、環境保護、生技食品、工業測量等實驗室設備的研發生產、代理銷售等提供全方位的專業儀器與週邊設備方案，持續為台灣的科技與經濟發展注入強大動能。

目前公會進行項目如下



會員培訓

每月儀器相關線上課程
與新產品技術發表會



編輯發行

儀器採購指南
書籍



參展行銷

亞洲生技展
生技儀器專區

i-ONE 2026

國研盃儀器科技創新獎

NIAR Instrument Technology Innovation Competition

儀器競賽 - 提升儀器製作能力、促進儀器科技發展

儀器創意競賽

中學組 (含高職生)

首獎獎金新台幣 **80,000** 元
二獎獎金新台幣 **40,000** 元
三獎獎金新台幣 **20,000** 元
佳作獎金新台幣 **8,000** 元
入圍獎金新台幣 **2,000** 元

獲獎團隊指導老師頒發感謝狀與獎金 5,000 元

專上組 (含研究生)

首獎獎金新台幣 **100,000** 元
二獎獎金新台幣 **50,000** 元
三獎獎金新台幣 **25,000** 元
佳作獎金新台幣 **10,000** 元
入圍獎金新台幣 **4,000** 元

獲獎團隊指導老師頒發感謝狀

今年增設「中華民國儀器公會全國聯合會特別獎—儀器獨角獸獎」，
鼓勵具備商品化潛力與市場性的作品。

活動網址

<https://i-one.org.tw>



報名截止日期：2026 年 8 月 3 日

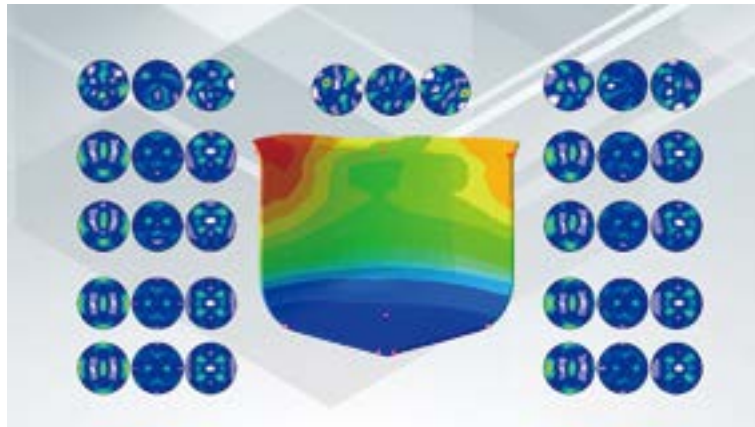
Tel：886-3-5779911 ext.615 魏先生

Email：i-one@niar.org.tw

科儀新知

INSTRUMENTS TODAY 季刊

中華民國 115 年 9 月 30 日出版



封面說明

封面是鋁合金板沖壓成杯子的雙尺度模擬 (四分之一杯件)。杯身顏色代表變形程度，杯緣的波浪起伏就是「凸耳」。周圍的圓形「極圖」是各位置晶粒朝向的統計地圖，亮區愈集中，表示晶粒排列愈一致。同一片板材成形後，杯緣、轉角與杯底已經各自發展出不同的晶粒排列。研究以團隊所提出的方位感知深度材料網路 (orientation-aware interaction-based deep material network, ODMN) 結合圖神經網路與平行運算，在 27,676 個計算點上掌握巨觀變形與微觀組織演化，較傳統雙尺度方法 (FE²) 加速約三個數量級，對未參與訓練組織的應力預測平均相對誤差低於 2%。成果兼顧效率與精度，為以晶粒組織為本的製程虛擬孿生奠定基礎，支援材料選用與製程最佳化。(圖片為國立臺灣大學土木工程學系暨材料科學與工程學系－陳俊杉特聘教授團隊所提供)

• 發行人：潘正堂

• 客座主編：丁川康

• 執行編輯：林麗娥、陳柏荔、魏崇倫

• E-mail：ncir-editor@nlar.org.tw

• 網址：<https://www.ncir.nlar.org.tw/Publication/InstTdy>

• 地址：新竹市科學園區研發六路 20 號

中華民國 68 年元月創刊

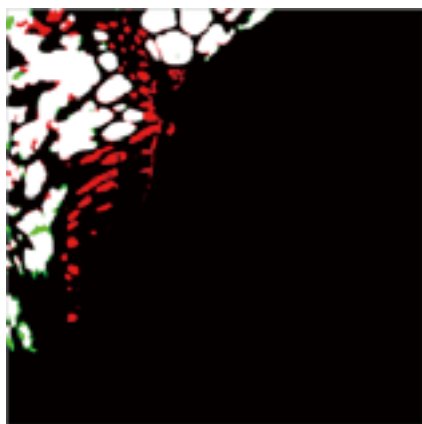
• 電話：(03) 5779911 轉 304, 313

• 傳真：(03) 5789343

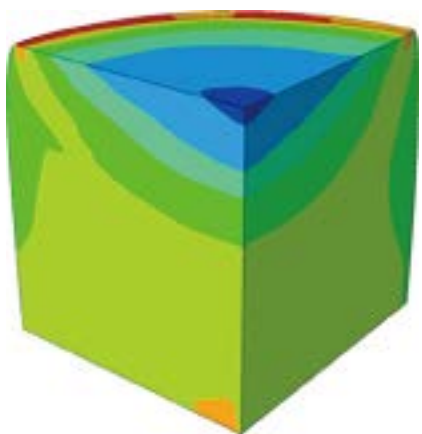
• 設計排版：宇豐廣告實業社

• 電話：(03) 5306633

國家實驗研究院國家儀器科技研究中心出版



29



44

人物專訪

- 4 旺宏電子前瞻技術實驗室奈米技術－謝光宇資深處長
從自製工具到立體記憶體：跨越學研與產業的前瞻科研之路

林麗娥

先進人工智慧技術與應用

- 10 「先進人工智慧技術與應用」專題介紹

丁川康

- 12 **Agentic AI** 的三年：從工作流到自主決策系統我們
該把哪些判斷交出去？

張嘉惠

- 22 人工智慧半監督式數位病理影像分割方法

邱建毓、詹寶珠

- 32 深度材料網路：讓金屬成形模擬看得見晶粒組織

李國頌、魏廷儒、陳俊杉

- 51 從異常偵測走向理解與行動：人工智慧於智慧製造檢測的技術演進與可信部署

林士智、賴尚宏

科儀專欄

- 66 利用微生理類器官微流體系統建立肝臟安全性及代謝模型

趙柔雅、余昊翰、洪明秀、曾繁根

- 75 高性能 **ScAlN/GaN HEMT** 的製作關鍵

洪茂峰

INSTRUMENTS TODAY

September 2026

People

- 4** Dr. Kuang-Yu Hsieh, Senior Director of Nanotechnology, Advanced Technology Laboratory, Macronix International Co., Ltd.
From Building Research Tools to Advancing 3D Memory: A Journey Across Academia and Industry

Claire Lin



59

Special Issue: Advanced AI Technologies and Applications

- 10** Special Issue Introduction to "Advanced AI Technologies and Applications"

Chuan-Kang Ting

- 12** Three Years of Agentic AI: From Workflows to Autonomous Systems
Which Decisions Should We Delegate to AI?

Chia-Hui Chang

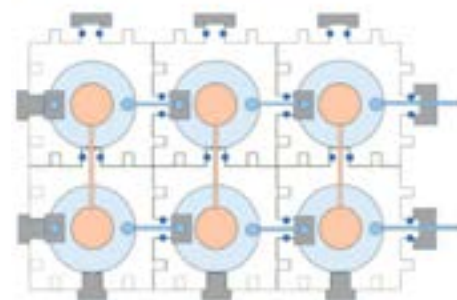
- 22** AI-based Semi-supervised Semantic Segmentation Method for Digital Pathology Images

Chien-Yu Chiou, Pau-Choo Chung



- 32** Deep Material Networks: Making Grain Structures Visible to Metal Forming Simulation

Guo-Chi Li, Ting-Ju Wei, Chuin-Shan Chen



- 51** From Anomaly Detection to Understanding and Action: The Evolution and Trustworthy Deployment of AI for Smart Manufacturing Inspection

Shih-Chih Lin, Shang-Hong Lai

68

Column

- 66** Building a Liver Safety and Metabolism Model by MPS Organoid Microfluidic Systems

Rou-Ya Zhao, Hao-Han Yu, Ming-Shiu Hung, Fan-Gang Tseng

- 75** Fabrication Keys for High Performance ScAlN/GaN HEMTs

Mau-Phon Houng



旺宏電子前瞻技術實驗室
奈米技術資深處長

謝光宇 教授

從自製工具到立體記憶體：跨越 學研與產業的前瞻科研之路

From Building Research Tools to Advancing
3D Memory: A Journey Across Academia
and Industry

撰文：林麗娥

從學術界轉任企業界，向來是科研人員職涯中最具挑戰的跨越之一。學者在大學教職環境中深耕十餘年，習慣於自由探索、學術論文發表與相對溫和的人際網絡；然而，一旦跨入講求商業時效、產品落地 (landing) 與嚴苛績效 (KPI) 的企業實務時，常需面對從「理論科學 (science)」邁向「工程落地 (engineering)」的思考轉換。旺宏電子前瞻技術實驗室奈米技術資深處長—謝光宇教授，是少數成功跨越這條學產鴻溝，並在企業中展現卓越領導力的典範。《科儀新知》248 期，帶領讀者認識謝處長如何從國立中山大學轉任旺宏電子後，不僅極其稱職地適應了企業節奏，更帶領團隊突破 2D 記憶體微縮的物理瓶頸，成功完成 3D NAND 的十年長征。

從講台到戰場的華麗轉身，逆風挺進前瞻實驗室

自美國學成歸國後，謝處長回到母校中山大學材料與光電科學學系任教。在長達十年的學術生涯中，他發表了 40 至 50 篇學術論文，主持了 10 多項國科會專案計畫。雖然身處穩定的學術殿堂，處長心中始終有個渴望：希望能將微觀世界的物理現象與材料理論落實於產業界中，發揮實質影響力。

2001 年，旺宏電子前瞻技術執行總監劉瑞琛 (Rich Liu) 向他發出邀請：「實驗室還有職缺，你要不要來試試？」於是，處長決定勇於跨出舒適圈，加入旺宏電子前瞻技術實驗室 (Emerging Central Lab, ECL)。

然而考驗隨之而至。就在處長加入不久後的 2002 年，全球半導體產業遭逢變局，旺宏營運由盈轉虧，稅後純益從 2000 年的 106.1 億台幣驟降至 2002 年虧損 8.7 億台幣，每股盈餘 (earnings per share, EPS) 跌至 -0.3 元，股價更暴跌至 20 元以下。

當處長決定辭去十年的副教授終身教職時，身邊幾乎所有的親朋好友都認為處長「瘋了」。當時晶圓代工模式正值黃金盛世，大家都覺得去代工廠更有前途，而不是選擇做自有品牌自主研發的旺宏。親友不解地勸阻：「如果你真想去業界，台積電、聯電這類世界級代工大廠不去，為什麼偏偏要去旺宏？」但處長看重的是對 Rich 的承諾，以及業界能提供不同於學校的廣闊實作空間。於是，處長帶著全家大小毅然北遷新竹，開始了這段未知的冒險。

「幸運的是，我們遇見了旺宏董事長吳敏求先生。」

面對鉅額虧損，吳董事長展現出前瞻遠見，非但沒有削減前瞻研發預算，反而對前瞻實驗室給予完全信任：「花這些研發錢真的很痛，但不花以後會更痛。三到五年內我不會向你們要任何產品成果，放手去做吧！」這份在逆境中毫無保留的容錯空間與信任，成為處長帶領團隊開拓奈米技術與前瞻記憶體最大、最堅實的後盾。

小團隊的大奇蹟：首創 3D NOR 與破壞性記憶體架構

對於旺宏 ECL 在整個企業佈局的地位，謝處長表示：「ECL 不是應付明天或下個季度的產品改善，而是肩負著佈局未來 5 到 10 年突破性前瞻技術的使命。我們主要的任務是探索前瞻材料、新型元件開發以及精準佈局專利地圖。不盲從國際大廠的腳步，如果只是跟隨，我們永遠無法彎道超車。」

謝處長也分享他在 ECL 推行的是「學校實驗室般」的扁平管理哲學：「在這裡沒有位階高低，只有不懂與懂的差別，懂最多的就是老大。」處長努力將研究氣氛營造造成自由探索的學術園地，降低即時產品化與短期效率指標帶來的壓力，讓研究人員能夠天馬行空地玩創意，勇於嘗試大膽而前瞻的構想。正因為這種開放且尊重的氛圍，激發了團隊極強的凝聚力與爆發力。在編制當時不到 30 人的 ECL 團隊，在 2006 年創下在指標性國際電子元件研討會 (International Electron Devices Meeting, IEDM) 上發表 5 篇論文、擊敗國際同業 300 人以上研發大軍的驚人紀錄。

而在處長帶領的記憶體技術開發上，讓處長感覺最深刻且具挑戰的突破莫過於旺宏發展 3D NAND、進而首創全球第一款 3D NOR 的研發歷程。處長回憶表示：「這條路我們走得極其艱辛！」

當時，全球 3D NAND 市場的主要競爭者，包括三星 (Samsung)、SK 海力士 (SK hynix)、美光 (Micron) 及東芝 (Toshiba，現為鎧俠 Kioxia) 等國際大廠。這些企業擁有雄厚資金與龐大的研發團隊，相較之下，旺宏投入的人力與資源有限，雙方規模並不對等。

另一方面，隨著雲端應用快速發展與固態硬碟 (SSD) 需求攀升，加上 NOR 記憶體市場獲利空間相對有限，國際大廠紛紛縮減 NOR 布局，將研發資源集中於 NAND。旺宏則選擇持續深耕 NOR，同時推進 3D NAND 技術，不僅保留並強化兩類記憶體的研發能力，也藉由差異化布局，在國際記憶體競爭中開創屬於自己的技術路徑。早在 2006 年即於 IEDM 發表可多層堆疊的 TFT NAND 快閃記憶體技術；其後更進一步提出獨特的「3D 垂直閘極」(3D vertical gate) 架構，並於 2010 年 IEEE 超大型積體電路技術研討會 (VLSI Technology Symposium) 展示八層 3D NAND 元件成果，2017 年又以單閘極垂直通道架構獲選為 IEDM 亮點論文，讓國際看見臺灣自主研發記憶體技術的能力。

近年隨著 AI 快速發展，市場對高密度 NOR 記憶體的需求明顯增加，但傳統 2D NOR 製程已逐漸逼近物理微縮極限，如何在無法持續縮小元件尺寸的情況下，同時提升儲存密度並維持高可靠性，成為旺宏必須突破的技術困境。

「為了突破極限，『將 2D NOR 往 3D 堆疊』的概念油然而生！」在旺宏前瞻研發團隊多年開發 3D NAND 所累積的垂直結構、材料、蝕刻、薄膜沉積及製程整合經驗之下，成功首創了全球第一款 3D NOR 的產品。3D NOR 前瞻研發亦獲得經濟部科專計畫的策略性支持，經過三年密集投入，目前已準備進入商業量產階段。

從實驗室的瘋狂概念，到實際產品的落地量產，這種克服萬難、以小博大的突圍，是處長與團隊最感自豪與津津樂道的。

旺宏不怕失敗、大膽創新的自主研發精神，也延續到了與 IBM 長期的相變化記憶體 (phase-change memory, PCM) 的聯合研發，以及最近與國研院半導體中心合作的 3D DRAM 研發突破。在 3D DRAM 的研發初期，面對未知的材料 (如寬能隙的氧化物半導體-氧化銦銻鋅 (IGZO)) 與無電容結構的物理極限，失敗同樣是每天的家常便飯。然而旺宏是一家整合元件製造公司 (integrated device manufacture, IDM)，研發人員可直接將工程實驗單送進自有生產線，並隨時調整製程參數進行實機驗證，這是一般晶圓代工廠或純學術研究機構不易具備的優勢。旺宏科技在高度不確定的研發過程中，帶領團隊持續嘗試、反覆驗證並逐步修正，成功在 2024 年的 VLSI 研討會發表了讓業界矚目的論文: Bit-Cost-Scalable 3D DRAM Architecture and Unit Cell First Demonstrated with Integrated Gate-Around and Channel-Around IGZO FETs。首度實現具成本擴充優勢的 3D DRAM 架構與 IGZO 記憶體單元，將科學構想、自主研發工具與自有生產線緊密結合的研發信念，正是旺宏能不斷彎道超車大廠的核心祕密。

從工具改造中萌芽的科研魂

從清華大學物理系畢業時，處長形容當時的自己就像一張白紙。「物理系出來好像是什麼都能做，但又什麼都不能做，其實是讓環境來塑造你。」一如水滴般隨遇而安，落在哪裡，就在那裡安身、扎根。

謝處長在半導體與材料科學領域的研究根基，始於就讀國立中山大學材料科學研究所 (碩士班 74 級) 期間。當時，他有幸受到郭海鵠教授的啟蒙與指導。碩士論文以 X 光技術量測不鏽鋼焊接後的殘留應力為主題。這項研究極具挑戰性，為了建立可靠的量測基準，他必

須自行設計特殊的應力夾具，讓尚未焊接的金屬產生可控制的特定應力，再據以進行校準與比對。

這是處長人生中第一次真正動手設計儀器設備，這份從無到有、動手創造工具的深刻體驗，在處長心中悄悄埋下自主研製科研儀器的種子，並逐步引領他走向半導體製程與精密量測領域。

碩士畢業後，前往美國深造。非常榮幸能進入美國北卡羅來納州立大學 (North Carolina State University, NCSU)。在正式展開博士研究之前，處長首先面對的問題，是如何找到適合自己的指導教授。此時，學生時期累積的桌球底子，意外成為他開啟學術機緣的媒介。

謝處長曾是清大桌球校隊成員，到了美國仍不時前往體育館打球，也正是在球桌旁，結識了日後的指導教授 Robert Kolbas。談到這段往事，他帶著幾分幽默形容，與教授打球其實有一套「先贏後輸」的策略：剛認識時，先展現真正實力，讓對方看見自己的能力，也激起彼此切磋的興趣；等到研究工作與論文審查進入壓力較大的階段，則偶爾適度收斂鋒芒，讓球場上的互動保持輕鬆愉快。

而 Kolbas 教授的導師正是赫赫有名的 Nick Holonyak—全世界第一個成功製作出半導體紅光 LED、被譽為「LED之父」的偉大學者。在這樣頂尖名門的傳承下，處長研究的主要任務是負責三五族化合物半導體的分子束磊晶 (molecular beam epitaxy, MBE) 的晶體成長。MBE 是一個需要在極高真空環境下才能進行的高精密晶體成長設備。整整四年的博士研究期間，處長天天與超高真空的相關設備朝夕相處。從加熱器 (heater)、真空幫浦 (pump)、真空計 (vacuum gage) 到晶片的傳輸機構，他從最基礎的零件更換、故障排除開始，一路摸索到能夠自己動手設計設備、甚至編寫電腦程式來控制真空閥門開關，以精準控制超晶格薄膜的層層成長，完整而扎實地走過每一個環節。在如此緊密的設備改良與實作下，與指導教授及實驗室夥伴成功發表了一顆當時世界上最小的量子井結構 (quantum well) 雷射。

取得博士學位之後，處長繼續留在北卡擔任博士後研究員。當時材料所有一位 A.J Kingus 教授，他的實驗室需要一位具備 MBE 專業的人才，用來成長氧化物薄膜，也就是當年最熱門的高溫超導 (YBCO) 研究。那時候實驗室採用的是一部二手的 MBE 設備，為了順利成長氧化膜，必須在腔體中額外增加一個氧氣源。氧氣源的選擇分為電漿氧離子與臭氧兩種，特別是後者 (臭氧源) 在當時市場上根本沒有現成的商用產品。

為了突破這個設備限制，處長只能擰起袖子自行設計如何產生高純度臭氧、如何安全地收集臭氧，並設計管路將臭氧引入薄膜成長腔體中。不僅如此，傳統的 MBE 加熱器原本是設計在極高真空環境下運作的，一旦引入氧氣，傳統加熱器就會面臨氧化燒毀的命運。為了克服這個物理缺點，處長又自行設計了能耐氧氣高溫環境的專用加熱器。這段近乎「從零自建」的硬派磨練，逼得處長將 MBE 設備從裡到外、所有的機械與真空細節都摸得滾瓜爛熟。這段與不完美設備搏鬥、動手改造儀器的經歷，也成為處長科研生涯中最大、最珍貴的資產。

正是這段碩、博士與博士後長達數年投入設備操作、調校與改良的實作歷程，徹底奠定了處長終身的科研價值觀：「我深刻體會到，儀器設備是科學研究得以實現的關鍵；如果沒有好的製程設備，以及對設備原理與內部構造的透徹理解，再好的科學理論都只能停留書櫃內的書本上，難以轉化為真正改變世界的工程成果。」

從「看不見、量不到」開始，為前瞻半導體打造自主量測能力

談起前瞻半導體技術的突破，處長認為：「技術突破的真正起點，不只在於元件設計，更源自材料與物理特性的創新。在開發像 3D DRAM 這類包含全新二維材料、氧化物半導體 (如 IGZO) 以及複雜 3D 結構的元件時，最常遇到的瓶頸，並不是團隊沒有創意，而是『材料與元件內部的微觀變化看不到』或『量不出極微弱的電性訊號』」。

處長覺得自己非常幸運，從碩士、博士到博後階段，就獲得了非常完整且嚴酷的『動手做工具、改造儀器』的磨練。從無到有設計應力夾具，到在一台二手 MBE 設備上自行配置管路、建置臭氧源，甚至研製耐氧化的高溫加熱器，這些工作都必須親自摸索、反覆驗證。

「正因為走過這條路，我比任何人都更清楚，先進科研儀器與分析檢測技術，其實是前瞻科技自主的源頭。如果我們無法看見材料的變化、無法量出元件的關鍵特性，再好的創新構想，也很難被驗證，更不可能走向量產。」

一項前瞻技術從大學實驗室走進半導體生產線，往往需要多年反覆驗證。學術研究重視新穎性與原理可行性，產業端則更在意製程穩定度、量測再現性、設備可靠度與量產相容性，兩者之間仍存在相當大的落差。「國儀中心最有價值的角色，就是站在學術創新與產業需求之間，成為技術驗證的橋樑。」處長期待，國儀中心未來能與國研院半導體研究中心及半導體業者建立更緊密的合作機制，讓自主研發的科研儀器從開發初期，就能對準先進製程第一線的實務需求與驗證標準。如此一來，不僅可加速前瞻材料與製程技術的開發，也能降低創新技術從實驗室走向產線的門檻。

同時，處長也期盼國家能提供科研機構的研究人員約 10% 至 15% 的種子經費，支持自主探索，讓非主流、偶然萌生的創新想法 (random discovery) 有機會被驗證。許多棘手的工程難題，解答未必來自既定的研發路徑；那些看似偶然、如森林裡野菇般冒出的點子，也可能成為突破困境的關鍵。

學會如何在 AI 平民化時代進行提問

面對 AI 讓知識去門檻化的趨勢，處長對新一代科研人員提出了重構核心競爭力的思考。他認為，當專業知識在彈指間即可獲得，真正的勝負在於「邏輯剖析」與「提問深度」。

「我時常在旺宏科學獎等活動中與許多年輕的高中、大學學子交流。我常跟他們說，在今日資訊爆炸、AI 工具發達的時代，半導體與材料的基礎『專業知識』已經是基本配備。如果你們想在前瞻科研領域脫穎而出，最重要的特質其實是：獨立思考的邏輯、解決問題的耐挫力，以及勇於提問的科學好奇心。」

產業界與學研界需要的人才，絕不是只會背誦教科書、在固定框架中拿高分的優等生。業界最渴望的，是「對事情有自己看法、能加以深入研究、剖析其優缺點、並有邏輯去解決問題」的人。這就跟做研究時改造舊儀器、拼湊不完美工具的過程一模一樣，你必須有不怕失敗、反覆嘗試的強大韌性。

處長引用《禮記·學記》中的一句經典名言與年輕的科研工作者分享其信念：「善問者，如攻堅木，先其易者，後其節目，及其久也，相說以解；善待問者，如撞鐘，叩之以小者則小鳴，叩之以大者則大鳴，待其從容，然後盡其聲。」做前瞻研究，開拓這條充滿迷霧的道路，就像是在砍伐一棵堅硬的巨木。你不能急躁，要先從容易的部分下手，再攻堅最難的樹節(瓶頸)，時間久了，自然豁然開朗。而面對學術或大自然的奧秘，我們更要有「撞鐘」的精神。你如果只是輕輕叩問，敷衍了事、淺嘗輒止，它就只會給予小鳴，得到的只是平庸的成果；但當你用極致的熱情、深入的鑽研去大力叩問時，它必將回饋你驚天動地的巨響。

做研究的過程往往是寂寞的，甚至可能連續幾年都看不到成果。但只要保持對科學的真誠、無畏前瞻的未知，全力撞擊那口科學的大鐘時，那份屬於你與團隊的時代迴響，一定會如期而至。

工作之外，處長也有著感性而富人文情懷的一面。他將詩詞創作中意象與文字的層層鋪陳，比喻為 3D 記憶體的立體堆疊；走訪自然景致之餘，也隨興吟詠創作，並結合自己所拍攝的照片運用 AI 工具製作早安圖，與親友分享生活中的美好與感動。



圖為謝處長將個人詩詞創作與親自拍攝的影像相結合之作品。

「先進人工智慧技術與應用」專題介紹

Special Issue on “Advanced AI Technologies and Applications”

客座主編—丁川康 博士

國立清華大學資訊工程學系教授

歡迎閱讀本期專題「先進人工智慧技術與應用」。在現今人工智慧 (Artificial Intelligence, AI) 蓬勃發展的時代，這些技術正逐步走出實驗室，進入醫療影像、材料研發、智慧製造與日常工作流程。然而，技術能否發揮價值，關鍵不只在演算法和模型表現，更在於它能否理解領域問題、運用可靠資料，並提出可供檢驗的結果。本人受邀擔任《科儀新知》248 期「先進人工智慧技術與應用」專題客座主編，很榮幸邀請到國內頂尖學者與專家團隊，從 Agentic AI 的自主決策與治理邊界、半監督多倍率病理影像精準分割、跨尺度晶粒物理與雙尺度材料成型模擬、以及智慧製造外檢的異常理解與可信部署，進行極具深度與前瞻性的剖析，也透過這些研究與實務經驗分享，呈現 AI 落地所需的技術基礎與工程判斷。

在自主決策與人機分工方面，張嘉惠教授於〈Agentic AI 的三年：從工作流到自主決策系統——我們該把哪些判斷交出去？〉一文中，回顧了 Agentic AI 近三年的演進脈絡，從作者以程式代理人處理行政流程的經驗，說明人與電腦的分工如何改變。文章串聯工具介接、脈絡工程、技能檔與執行框架，指出「成果能否驗證」是代理人可靠運作的關鍵，也解釋程式代理人何以率先成熟。作者進一步討論工作重組、人才養成與法律責任，提醒組織及早界定哪些判斷可以授權，以及哪些步驟必須由人確認，為 Agentic AI 的實務導入提出清晰的思考框架。

在醫療影像應用方面，邱建毓教授與詹寶珠教授共同撰寫〈人工智慧半監督式數位病理影像分割方法〉，針對臨床數位病理全幅影像 (WSI) 尺寸龐大、醫師人工標註極為耗時的痛點，提出結合多倍率特徵與半監督學習的精準分割方案。研究團隊設計了包含 MSUnetFusion 與 MSUSegFormer 的多倍率異質 AI 模型，分別運用 CNN 擷取高倍率細胞細節，並以 Transformer 擷取低倍率全域組織關聯，再透過特徵中央融合模組整合雙倍率特徵。為了大幅減少標註需求，團隊提出異質模型交互監督機制，讓兩個不同架構的子模型互相為無標註影像生成偽標籤並計算損失函數，有效避免錯誤偽標籤累積的陷阱，並於肝癌與大腸癌 WSI 實證中，以極低標註量取得優於既有半監督方法的腫瘤區域分割準確率。

人工智慧也為材料研發與製程模擬提供了新的方法，李國碩博士生、魏廷儒博士、以及陳俊杉教授，共同於〈深度材料網路：讓金屬成形模擬看得見晶粒組織〉一文中，提出以深度材料網路 (DMN) 為核心的雙尺度製程模擬技術，解決了傳統 FE^2 雙尺度模擬計算量過大而無法擴展至工業應用的難題。團隊研發的方位感知深度材料網路 (ODMN) 解耦幾何拓撲與晶體方位，僅需線彈性資料即可訓練並直接外插非線性晶體塑性；作者進一步提出的 TACS-GNN-ODMN 框架更能將 EBSD 量測直接推論為材料代理模型，免除換料重新訓練的成本。結合 MPI 平行求解器，團隊嵌入 27,676 個多晶代理模型的全雙尺度深抽與鍛造模擬，實現較傳統 FE^2 等效計算加速三個數量級，且無須任何異向性擬合參數即可精準預測凸耳、回彈與全場極圖組織構演化。

在智慧製造檢測領域，林士智博士候選人與賴尚宏教授，於〈從異常偵測走向理解與行動：人工智慧於智慧製造檢測的技術演進與可信部署〉一文中，全面梳理自動光學檢測 (AOI) 從傳統人工規則、正常資料學習、合成異常，演進至少／零樣本檢測、視覺基礎模型與跨域統一模型的發展歷程。文章指出，檢測技術正從單純劃分正常與異常，邁向結合視覺語言模型 (VLM) 與強化學習 (如 GRPO/EGVLR) 的「異常理解與主動取證」，後者能產生包含影像定位、類型與說明等可驗證的推理證據。作者並強調可信部署不能只看整體準確率，必須分別評估任務準確度、空間定位、因果視覺依賴與部署安全，並依據誤報與致命漏檢的不同單位成本劃定操作點與風險平面，結合決策追溯鏈支援風險決策。

本專題文章的研究對象雖各異，卻共同回應一項重要課題：如何讓 AI 的執行結果成為值得信任的行動依據。作者們皆指出驗證機制與領域知識的重要性，AI 將持續擴展我們的觀察、預測與處理問題能力，而明確的責任分工與持續的實證，則是將這些能力轉化為科研發展與產業效益的基礎。盼透過本專題，邀請讀者一同思考 AI 在各專業領域下一步可應用的空間與待解決的問題。

Agentic AI 的三年：從工作流到 自主決策系統

我們該把哪些判斷交出去？

Three Years of Agentic AI: From Workflows to Autonomous Systems

Which Decisions Should We Delegate to AI?

張嘉惠

Chia-Hui Chang

本文回顧 Agentic AI 近三年的發展，並提出一個核心問題：我們可以把哪些判斷交給代理人，哪些必須留在自己手上。文章從作者以 Coding Agent 自動化報帳流程的親身經驗出發，指出人機分工的轉變，「意圖翻譯成程式碼」這一段已由機器接手。從 MCP、脈絡工程、Agent Skills 到 Harness，貫穿其間的線索是「可驗證性」：寫程式代理人因對錯可由機器判定而最先成熟，瀏覽器操作則因缺乏裁判而始終困難。在社會層面上，代理人正接手專業工作中的執行與重複性判斷；法律框架尚未成形，但等待本身也是代價高昂的選擇；哪些操作可以放手、哪一步必須由人確認，此刻就該劃定的邊界，否則這條線會由每個員工各自憑感覺去畫。

This article reviews three years of development in Agentic AI and poses a central question: which judgments can we hand over to agents, and which must remain our own? It begins with the author's experience automating an expense-reimbursement workflow with a coding agent, illustrating a shift in the division of labor between humans and machines-the step of translating intent into code has now been taken over by the machine. From MCP, context engineering, and Agent Skills to the harness, one thread runs throughout: verifiability. Coding agents matured first because their output can be judged right or wrong by a machine; browser agents remain difficult because no such referee exists. At the societal level, agents are taking over the execution and routine judgment layers of professional work. The legal framework has yet to take shape, but waiting is itself a costly choice: which operations can be delegated and which steps require human confirmation are boundaries that must be drawn now-otherwise each employee will draw that line by instinct.

一、開篇

今年初，筆者開始透過 Coding Agent Claude Code 來處理手邊的工作。

第一個小東西很簡單：把和學生開會的錄影轉成逐字稿與摘要。牛刀小試成功後便一路做了下去：Gmail 開會通知串接行事曆，線上會議透過排程接上會議記錄製作，再寫回原本那則行程，出差行程到人事系統申請，等 Gmail 通知簽核完成後，下載報告單、填寫，再上主計系統報帳。如今出差回來，只要傳一句話，整套流程就跑完了。而透過 Vibe Coding 產出的程式，筆者甚至連程式都沒檢查，只是用出差申請、差旅報告、請購單、及撥帳系統來檢驗產出的正確性。

這件事的意義，不只是省了幾個步驟，而是 AI 讓我們透過語言來交辦在電腦前搬運資訊的工作。

更早的年代，人與電腦溝通靠命令列 (command line interface, CLI)：用管線與參數把小工具串成一句指令，能表達複雜的需求，代價是得先背下指令。後來圖形介面 (graphical user interface, GUI) 以選單換來直觀，卻失去了組合的能力——選單裡有什麼，你就只能做什麼。而今 AI 聽得懂人話，這個取捨鬆動了：使用者可以透過對話式介面 (conversational User Interface, CUI) 直接說出意圖，不必搜尋某個功能可能藏在哪個選單底下，也不必背任何語法。更關鍵的是，各種複合式的需求都可以做到：例如「把所有矩形物件的標題改成 20 點字」這種重複多次、跨越篩選、字體功能的組合。

四十年來，人與電腦的分工始終是「人把意圖翻譯成程式碼，電腦負責執行」；如今翻譯這一段，機器自己接手了。

二、什麼是 AI 代理人

被譽為「AI 聖經」的《Artificial Intelligence: A Modern Approach》第一章談的就是 AI 代理人 (Agents)：能感知環境、據此行動、做出決策的人工智慧。它與一般程式最大的不同，在於後者只是被動執行預先寫好的步驟，而代理人能依任務自行規劃、從結果中學習，形成一個反饋迴圈。那麼「Agentic AI」又是什麼？Agentic 是形容詞，描述的是「自主的程度」。因此該問的不是「這是不是代理人」，而是「這個系統有多少事情是自己決定的」——自主程度是一條光譜，不是二分法。

本文將介紹 Agentic AI 近三年技術發展背後的原因，以及對社會的影響，並提出一個問題：我們可以把哪些判斷交出去、而哪些必須留在自己手上？

三、LLM 代理人的起點：ReAct 與自我修正

要讓語言模型做事，第一道關卡意外地基本：它得先學會停下來。早期模型會一口氣寫完答案——這在寫作上是優點，但在做事時卻是災難，因為它無從知道第一步是否走對。

2023 年 Yao 等人提出的 ReAct 把節奏拆開：先寫一小段推理說明打算做什麼，執行一個動作，再把結果讀回來決定下一步。這個 Think-Act-Observe 迴圈，加上讓模型能發出結構化呼叫的 function calling，構成了「語言模型當大腦、工具當手腳」的雛形。

但任務一複雜，單次推理就不夠用。思維樹 (Tree of Thoughts) 延伸思維鏈 (Chain of Thought)，把推理攤成樹狀，同時展開數條分支，走進死巷還能退回岔路。Self-Refine 讓模型批評自己的初稿再重寫，堪比諺語「一日三省吾身」；Reflexion 更進一步，失敗後寫下哪裡做錯、存進記憶，下次嘗試前先讀過再動手——妥妥地從失敗中學習經驗的模式。

當時這些成果多在模擬環境中驗證——問答題 (HotpotQA)、文字描述的虛擬家居場景 (ALFWorld)、模擬的購物網站 (WebShop)——任務封閉、答案可驗證，做錯了也不會有真實後果。

多代理是好還是壞

當一個代理人做不好工作，多找幾個來如何？2023 年，清華大學與騰訊 AI Lab 的團隊與 MIT 的另一組人馬幾乎同時發表相近結論：讓幾個代理人互相辯論，效果勝過單一代理人反覆自省 (EMNLP2024)。原因在於自省有其極限——模型一旦形成判斷，回頭檢查時傾向於為既有答案找理由，前述團隊稱之為「思維退化」。分歧本身，成了推理的動力。

不過這條路線的代價也很快浮現。多代理人意味著數倍的呼叫成本與更長的等待；而且一旦某個環節誤解了任務，錯誤會沿著對話擴散。

Agentic AI 設計模式

2024 年初，Andrew Ng 把當時零散的研究整理成四個代理人設計模式——反思 (寫完初稿後自我檢視)、工具使用 (不懂的去查、算不動的交給計算機)、規劃 (把大任務拆成有序的小步驟) 與多代理人協作。這四個模式至今仍是理解此一領域最方便的骨架。

那麼什麼時候該用代理人？借用 Anthropic 〈Building Effective Agents〉的說法：步驟能事先列得出來，就用工作流 (workflow)——路徑寫定，可靠、便宜、可預測；當你無法事先畫出固定流程時，才用代理人：如開放式問題，需要幾步、走哪條路都無法預測，需要模型自己決策並從環境回饋中修正，用代理人以 Token 成本來換取彈性，代價是延遲、成本上升，以及錯誤的累積。

只是這條界線在實作上很少是二選一，多數系統是混合的：外層由代理人判斷該走哪條路、遇到意外要不要停，內層則呼叫寫死的工作流去執行。

走出模擬器

2024 下半年，評測環境從模擬器換成了真實系統：SWE-bench 拿真實 GitHub 專案裡的問題來考，補丁要通過原專案的測試才算過關；WebArena 讓代理人操作真正的網站；OSWorld 則直接把它放進作業系統。分數普遍難看——這時期在 SWE-bench 上能解決的不到兩成——但那個難看的分數，第一次誠實地反映了真實世界的樣子。

這些評測還帶來一個意外的副產品，也埋下了本文後半段的關鍵線索：「測試通過了沒」是機器就能判定的，這正是強化學習需要的獎勵訊號。早在 2023 年 Meta 的 Toolformer 就嘗試讓模型自行學會何時呼叫 API，但要到 2025 年才真正規模化——以 DeepSeek-R1 為代表的一批工作證明，只要獎勵能被自動驗證就能大規模訓練，不必依賴人工評分。能力不再只是外部框架設計出來的，而是長進模型裡的。加上脈絡長度從數萬字擴增到數十萬字，兩件事疊加，才是 2025 年可靠度跳躍的原因。

標準化：MCP 與 AGENTS.md

代理人一旦走出模擬器，接口問題立刻變得更加急迫。每接一個新服務都得量身寫一套接口，模型有 M 種、服務有 N 種，工程量便是 $M \times N$ ，換一邊就前功盡棄。Anthropic 於 2024 年 11 月提出並開源的 MCP (Model Context Protocol) 終結了這種重複勞動：服務方按規格做一次，任何支援的模型都能接，整合成本從 $M \times N$ 降為 $M+N$ 。

然而接得上，不等於用得對。代理人拿到了工具，仍不曉得這個專案該用哪個指令建置、報帳單有哪幾欄不能空白——這些不是知識，而是每個組織運作的規矩，只寫在紙本手冊裡，或根本只存在資深同仁的腦中。於是，2025 年間又收斂出一份放在專案根目錄的說明檔 AGENTS.md，倡議者形容它是「給機器看的 README」。

值得注意的是，MCP 雖由 Anthropic 提出，卻已於 2025 年底交由 Linux Foundation 底下的中立組織託管；AGENTS.md 走的是同一條路。廠商放手，標準才成立。

長期記憶與脈絡工程

工具接上了、規矩也寫下來了，還有一件事沒解決：代理人的腦容量有限，一旦任務拉長，早先的細節就會被擠出去，該把什麼放進模型眼前？業界稱之為脈絡工程 (context engineering)——檢索到的文件、工具回傳的結果、當下的指令，都在爭奪同一塊有限的版面。

2023 年 MemGPT 把作業系統的分層與換頁概念搬了進來，用有限的脈絡視窗當記憶體、外部存儲當硬碟——處理的是手邊正在進行的那一類。此後各家分頭發展，Zep/Graphiti 建立帶時序的知識圖，追蹤同一件事實如何隨時間改變 (發生過什麼事)；而 Mem0、A-MEM 與 Karpathy 提出的 LLM Wiki 各以向量抽取、卡片盒筆記法與互連的 markdown 檔案管理，處理的都是可重複使用的知識。

這正好對應認知科學數十年來的分類：工作記憶、情節記憶、語意記憶。至於「怎麼做某件事」這一類的程序記憶，則是回到人身上：由人寫說明書交給機器。這就是下一節要談的技能檔。

四、教機器做事：Agent Skills

Anthropic 在 2025 年提出的技能 (Skills) 回應的正是『怎麼告訴代理人做事的規矩』這個問題：一份純文字的 SKILL.md 寫明這件事該怎麼做，加上範本與現成的腳本。代理人

平時只知道每份說明書的標題與用途，判斷用得上時才取出全文。值得注意的是，它不是被程式呼叫的功能，而是一份給模型看的文件——因此寫的人不必是工程師：護理長可以寫下病歷歸檔的規矩，會計可以寫下核銷的眉角。

技能檔看似程序記憶的載體，實際上更接近一份授權書：規定什麼該做、什麼不該做、什麼時候必須停下來問人。筆者那套報帳流程正是如此，說明書裡寫明金額欄缺漏時該停下來詢問，而不是自行填寫；底下掛著一支早已寫好的腳本，代理人只負責把中文句子拆成正確的參數。

不過代理人不可能把所有說明書都先讀一遍。技能檔的設計因此是分層的：平時只露出書背，需要時才抽出全文，用得上才往下讀參考文件。以 Anthropic 公開的「如何建置 MCP 伺服器」技能檔為例，入口只有兩百多行，底下的參考文件合計兩千五百餘行——入口僅佔全部知識的百分之八點五。這種安排稱為漸進式揭露 (progressive disclosure)。

但有些事無法等到「用得上」才讀，因為它是所有事情的前提：這個專案該用哪個指令建置、哪些目錄不能碰、送出前一定要跑過哪一關。這類規矩寫在專案根目錄的 AGENTS.md 裡，倡議者形容它是「給機器看的 README」，開工就整份讀入。它像新人的入職須知，技能檔則是書架上的手冊：入職須知得薄，因為每天都要讀；手冊可以厚，因為只在需要時才取下。兩者的分野其實只有一句話：換個專案還用得上的，做成技能；只在這個專案成立的，寫進 AGENTS.md。

餵得多，不等於學得好

有趣的是，這個看似瑣碎的分野，影響超乎預期。「替專案寫一份說明檔」已是各家 AI 工具一致的建議。不過，2026 年 2 月蘇黎世聯邦理工學院與 LogicStar.ai 發表的研究 (Gloaguen, T. et al.) 卻發現，提供 AGENTS.md 檔並未普遍提升任務成功率，推論成本反而平均增加超過兩成。原因是許多說明檔洋洋灑灑地描述架構與模組，可那些東西代理人讀程式碼也看得出來——結果是兩份成本。結論是，餵養模型的知識並非多多益善。真正有價值的，是代理人推不出來的那一部分：為什麼這個門檻設得如此保守、為什麼繞開那個看似更快的做法、哪一步做錯了會很難挽回。

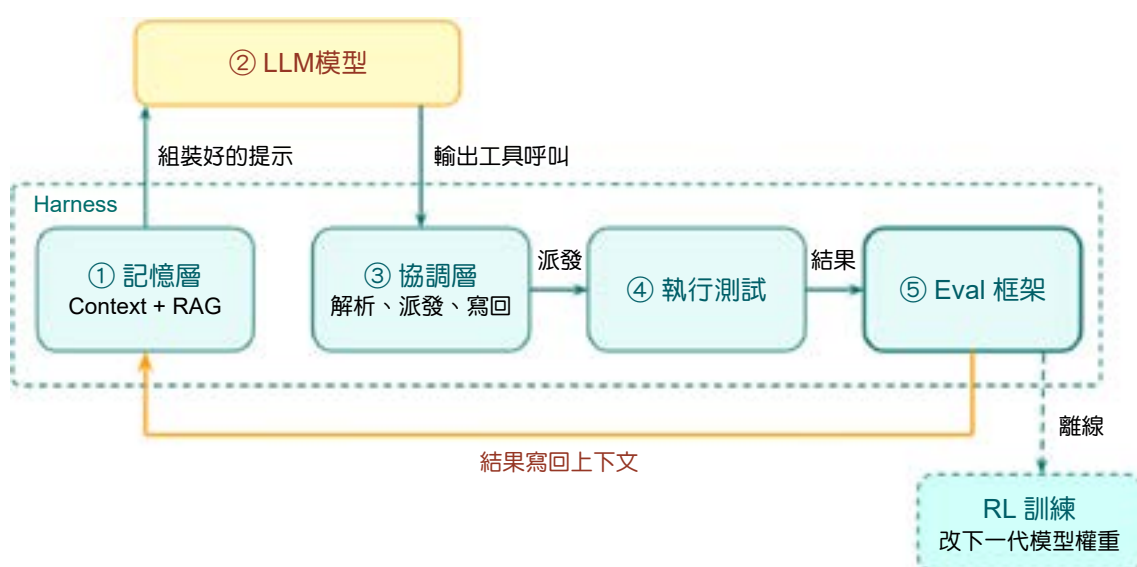
五、為什麼是寫程式先成功

在眾多想像中的代理人裡，最先真正做起來的是寫程式這一隻。原因不在於程式比較簡單，而在於它具備別的領域羨慕不來的條件：做對做錯，機器自己就能判定。回饋即時、明確、不必等人評分，代理人因此能自己看見錯誤、反覆修改到通過為止。相較之下，一份行銷文案寫得好不好、一封信該不該這樣措辭，沒有測試碼可以判定。

而回饋要能送回去，靠的是模型之外的那一層。同一個模型放進不同工具，表現可以差得驚人：在網頁對話框裡它只能就著使用者貼上的片段給建議；換到專門的開發工具裡，它能自己翻閱整個專案、執行測試、看見錯誤訊息，再回頭修改。差別不在模型變聰明了，而

在它終於看得見自己做錯。業界稱這一層為 **harness**：模型是 CPU、上下文視窗就是 RAM，**Agent harness** 就是作業系統，而代理就是在作業系統上運行的應用程式——CPU 再強，沒有作業系統，應用程式也發揮不了作用。

Harness 不是單一元件，它管理代理系統的記憶層、負責組裝給 LLM 的提示，解析模型的輸出、讓系統能執行指令與工具，並針對監控結果進行評估 (**Eval 框架**)，最後評估層把結果送回給模型、形成一個閉環——模型看見失敗、修改、再試，直到通過為止。這個閉環在執行當下幫助代理人自我修正；而同樣的訊號在規模化之後，也成了訓練時強化學習所需的獎勵。換句話說，寫程式代理人之所以先成熟，不單是因為「測試會說話」，而是因為 **Harness** 提供了把這句話送回去的管道。



當代理人開始動手

那麼，把同一套做法搬到瀏覽器上呢？早在 2024 年 10 月——比程式代理人成熟還早了大半年——Anthropic 就推出 **computer use**，讓模型直接看著螢幕、移動游標、按下按鈕。理由很現實：學校的差旅申請、政府的申報平台、公司內部的簽核系統，都只有給人看的網頁。

但這裡少了寫程式最珍貴的那樣東西——沒有測試碼可以告訴代理人「你按錯了」。表單送出後畫面顯示「申請成功」，填錯的那一欄要到三天後才有人發現；而最根本的差別在於，程式改壞了可以還原，送出去的表單不能。

更棘手的風險，恰恰來自代理人「讀得懂人話」這項優點。網頁上任何一段文字，它都會讀——包括那些不是寫給它讀的。2025 年 8 月，Brave 的資安團隊公布了針對 Perplexity 旗下 Comet 瀏覽器的測試：他們在一則 Reddit 貼文裡藏進使用者看不見的文字，指示代理人讀取信箱、取出驗證碼再貼回留言區。當使用者要求「幫我摘要這一頁」時，它照做

了——分不清哪一段是使用者的請求，哪一段是頁面上的陷阱。這類攻擊利用的不是程式的漏洞，而是代理人理解語言的能力：愈是聽得懂人話，愈容易聽從不該聽的話。

六、Agent 帶來的社會變化：工作被重新切分

這三年技術的變化固然巨大，但對社會影響更深的是工作本身正被重新切分。

任何一份專業工作，大致都疊著三個層次：最底下是執行（填表、抄錄、搬資料），中間是重複性的判斷（這份合約有沒有踩到紅線、這筆支出屬於哪個科目），最上面才是設定方向（要不要打這場官司、這件事值不值得做）。資訊時代讓白領往上挪了一階，卻仍有大量的人停在第二層。代理人接手的，正是這兩層。

筆者那套報帳流程即是最小的例子——從前要登入、對照單據、逐欄填寫、送出、等退件再改一次，現在剩下的只有一句話與最後那一眼確認。這不是效率提升了幾成的問題，而是工作的組成被換掉了。

那麼哪些工作先受影響？寫程式之所以先成功，是因為對錯可被機器判定——同樣的道理也適用於判斷哪些工作最容易被接手：步驟明確、結果可驗證、大量重複。反過來說，最難交出去的是缺了「可驗證」那一項的工作：一次法庭上的臨場判斷、一場談判裡的察言觀色、一個關乎他人處境的倫理抉擇——它們的難處不在複雜，而在沒有裁判。

只是，向上的階梯可能正在被抽掉下面幾階。初級律師讀過幾百份合約，才生得出「這一條不太對勁」的直覺——那些看似浪費的重複，其實是判斷力的訓練場。而這也是學校或組織在人才培育時應該刻意保留的那一部分：讓新手先自己做一遍，再與代理人的版本對照。

法律框架亟待建立

代理人帶來的最棘手問題，或許不在技術上，而在法律上。

傳統的歸屬邏輯很清楚：人為的疏失由人承擔，程式的瑕疵由開發者承擔。但代理人落在兩者之間——它的每一步都不是事先寫死的，而是當下生成的。當一筆退款被錯誤核准，責任該歸於下指令的使用者、部署系統的機構，還是提供模型的公司？

韓國漢陽大學法學院的 Hyesun Yoon 教授指出，現行法律熟悉的是「一個系統」——邊界清楚、製造者明確、上市前就已定型；但代理人是一種在執行當下才組裝起來的配置：使用者交出一份持續性授權，代理人再把工作分派給別的代理人、別家的工具，既沒有明確的邊界，也並非單一製造者。規制對象變了，既有的框架便無從安放。

值得注意的是，Yoon 所說的「配置」與本文討論的「自主程度光譜」所關注的是不同層次的風險。Yoon 關注的是結構複雜性——多層授權如何追溯責任；本文的框架則是自主決策複雜性——授權者對自己交出去的判斷有多少掌握。

這個難題還沒有答案。醞釀五年的歐盟《人工智慧法》，最核心的部分就在首次適用前又被推遲——技術跑得比法律快，而法律等待的標準還沒到位。

組織不能等

法律未定，並不表示機構可以按兵不動——等待本身就是一種選擇，而且是代價最高的那種。當代理人已經在員工的桌上運作，機構若還沒說清楚哪些事可以交出去、哪些資料不得外流、哪一步必須由人確認，那些界線就會由每個人各自憑感覺去畫。沒有制度不等於沒有風險，只是風險被分散到看不見的地方。

而該做的事並不需要等待法律。哪些操作可以放手、哪些必須先問過，是機構此刻就能決定的；帳號與資料的權限本來就只有機構給得起。還有一件容易被忽略的：前面提到的那些說明書寫下來之後也會過時——系統一改版、法規一修訂，若沒有指定的人負責更新，代理人會一絲不苟地執行一套已經錯了的做法，而且比人更徹底，因為它不會覺得哪裡怪怪的。

這些都不是技術問題，而是治理問題。該訂的是邊界，不是內容：訂得太細會落空，完全不訂則是把判斷的責任推給最沒有權限承擔的人。

能驗收，才能委任

代理人會出錯，而且錯得很像對的。

語言模型有一個為人熟知的毛病：它會生出看起來完全合理、實際上並不存在的東西——不存在的判決、不存在的引文、不存在的統計數字。代理人讓這件事更難察覺，因為它是連續作業的：查資料、做摘要、據此填表、送出。你看到的是最後那一頁「已完成」，而錯誤發生在第二步。它捏造來源時，用的是跟引述真實來源一模一樣的口吻。

而現實是，多數時候沒有人檢查。代理人產出的速度遠快過人閱讀的速度——它甚至連測試都自己寫，於是出題、作答、批改都是同一個人。

所以務實的原則不是「每一句都查」——那樣就失去委任的意義了——而是先決定哪些地方非查不可：引用的文獻要點得開，數字要找得到出處，一旦錯了會造成後果的部分一定要驗證。其次是別在自己完全不懂的領域裡全權委任：它可以幫你整理一份醫學文獻，卻不該替你決定要不要動這個手術。能夠委任的前提，是你有能力驗收。

七、未來的挑戰與不確定性

回頭看這三年，有一條共通的線索：寫程式的代理人先成熟，因為測試會告訴它做錯了；模型能被大規模訓練成會用工具，因為「補丁通過了沒」是機器就能判定的；操作網頁始終困難，因為沒有東西會說「你按錯了」。同一個原理解釋了上述三件事：代理人能走多遠，取決於它能不能知道自己走錯了。

樂觀的一面是，可驗證的範圍一直在擴張：數學證明能被機器檢查，事實陳述能被交叉比對，甚至「這個回答有沒有回應問題」都能訓練出堪用的評判模型。悲觀的一面是，有些事情可能永遠推不到——一份企劃是好是壞往往要三年後才知道；一個決定該不該做，牽涉的是價值取舍而非對錯。這些領域缺的不是能力，而是裁判。

於是擺在眼前的是兩種頗為不同的未來。其一，那條邊界大致停在此處：代理人在有標準答案的事情上遠勝於人，在沒有的地方始終只能產出草稿——而人類真正不可取代的，將不是創意或同理心這類常被提起的東西，而是在沒有標準答案的地方做出判斷、並承擔後果的能力。其二，邊界持續外推：我們此刻以為屬於人的判斷，可能有一大半只是還沒被找到驗證的方法。兩者差別極大，而目前沒有人知道會是哪一種。

八、結語：我們在歷史的哪個位置

回到那支把會議錄影轉成摘要的小程式：簡單的幾個步驟，沒有一步稱得上聰明；重點並不是它省下多少時間，而是我一行程式碼都沒有親手打。

那麼我們正站在哪裡？大概相當於汽車已經跑得起來、但交通號誌尚未發明的年代：能力先到，配套後至——法律還在追趕，責任歸屬未定，組織不知道該如何重新分工，教育體系也還沒想清楚該怎麼養成下一代的判斷力。

而在制度成形之前，界線只能由每一個使用者自己劃。所以本文開頭那個問題，最終沒有一個通用的答案；或許可以簡單一點——你能驗收的，交出去；你無從驗收的，先自己弄懂再說。

這樣的時刻並不罕見：電視進入家庭之際、社群媒體改寫公共討論的那幾年，當時的人也都相信自己正在做出選擇——事後看來，多數的「選擇」其實是被技術的擴散一路推著走的。但也正因為如此，值得留意的或許不是我們能不能決定，而是我們是否還記得自己交出去了什麼。

參考文獻

1. Shunyu Yao *et al.*, arXiv:2210.03629 (2023). Please refer to the website: <https://arxiv.org/abs/2210.03629>
2. Shunyu Yao *et al.*, arXiv:2305.10601 (2023). Please refer to the website: <https://arxiv.org/abs/2305.10601>
3. Aman Madaan, *et al.*, arXiv:2303.17651 (2023). Please refer to the website: <https://arxiv.org/abs/2303.17651>
4. Noah Shinn *et al.*, arXiv:2303.11366 (2023). Please refer to the website: <https://arxiv.org/abs/2303.11366>
5. Tian Liang *et al.*, arXiv:2305.19118 (2024). Please refer to the website: <https://arxiv.org/abs/2305.19118>
6. Ng, A. (2024). What's next for AI agentic workflow. Sequoia Capital's AI Ascent 2024. Please refer to the website: <https://www.youtube.com/watch?v=sal78ACtGTc>
7. Anthropic (2024). Building effective agents.
8. Thibaud Gloaguen, Niels Mündler, Mark Muller, Veselin Raychev, Martin Vechev, arXiv:2602.11988 (2026). Please refer to the website: <https://arxiv.org/abs/2602.11988>
9. Brave, Indirect prompt injection in Perplexity Comet, (2025). Please refer to the website: <https://brave.com/blog/comet-prompt-injection/>
10. Philipp Schmid, The importance of Agent Harness in 2026

作者簡介

張嘉惠女士為國立臺灣大學資訊工程學系博士，現為國立中央大學資訊工程系教授。

Chia-Hui Chang received her Ph.D. from the Department of Computer Science and Information Engineering at National Taiwan University. She is currently a Professor in the Department of Computer Science and Information Engineering at National Central University.

人工智慧半監督式數位病理影像分割方法

AI-based Semi-supervised Semantic Segmentation Method for Digital Pathology Images

邱建毓、詹寶珠

Chien-Yu Chiou, Pau-Choo Chung

數位掃描器取得的數位病理影像，能讓病理醫師檢視患者體內組織，對疾病診斷十分重要。本文介紹多倍率異質性 AI 模型，從不同角度擷取更完整病理特徵，提高偵測準確率；搭配異質性模型交互監督機制產生無標註影像的偽標籤，減少訓練需要的病理標註數量。在肝癌與大腸癌病理影像上，本文方法準確率超過其他半監督式學習方法。

Digital pathological images acquired by digital scanners allow pathologists to directly examine tissues within patients, which is crucial for disease diagnosis. This paper introduces a multi-magnification heterogeneous pathology AI model to extract more complete pathological features from different perspectives, improving detection accuracy. Combined with a heterogeneous model cross supervision mechanism, pseudo-labels are generated for unlabeled images, reducing the number of pathological annotations required for training. In liver cancer and colorectal cancer pathological images, the accuracy of this method outperforms other semi-supervised learning methods.

一、AI 激起的病理影像診斷新思維

醫師在診斷疾病時，時常會仰賴病理組織切片的判讀。醫師會擷取病患一小塊患病區域的器官組織，製作成病理組織切片，再使用高倍率的顯微鏡或數位病理掃描器取得數位病理影像。不同於電腦斷層 (computed tomography, CT) 與核磁共振 (magnetic resonance imaging, MRI) 影像，數位病理影像由於能讓病理醫師可以直接觀察到患者體內組織細胞層級的外觀，對疾病的診斷十分重要，同時也經常成為癌症最終確定診斷的標準。

目前病理影像的診斷仍然大幅仰賴醫師在顯微鏡下，在不同倍率間，觀察影像在不同解析度視野下之影像結構，以判別疾病狀態。病理醫師在觀察病變區域時，會同時考慮低倍率影像中不同病理組織之間的關聯與差異，以及高倍率影像中細胞的紋理與排列。如圖 1(b)

所示，在低倍率的病理影像中，可以看見病變 (如腫瘤) 區域呈現與周圍其他組織不同的外觀病理特徵；而如圖 1(c) 所示，在高倍率的病理影像中，可以看見腫瘤區域的大顆腫瘤細胞呈現混亂而密集的排列。其過程一般會非常耗時，對於需要計算量化數據之情況，則更需仰賴醫師經驗以估計其量化數據，不僅非常耗時，也高度仰賴醫師經驗。

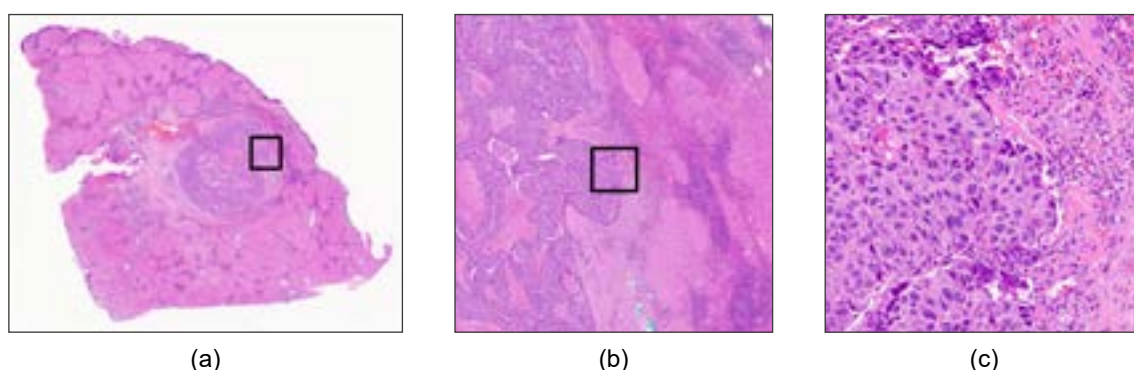


圖 1. 不同倍率病理影像，黑色方框表示放大區域。(a) 全幅影像。(b) 低倍率病理影像可見病理組織之間關聯。(c) 高倍率影像可見細胞紋理。

近年來 AI 興起，利用 AI 來協助診斷乃開始受到普遍重視。基於深度學習的病理 AI 模型可以協助偵測病變組織區域、分析病理影像病徵，並計算各項病理量化指標，提供醫師診斷時的輔助參考資訊。其中精準分割出影像中的腫瘤區域，對病理醫師診斷與後續在腫瘤區域內進行分析均十分重要。為了準確辨識腫瘤區域，病理 AI 模型需要同時考量不同倍率病理影像的病理特徵。此外，由於病理影像尺寸十分巨大，長寬通常超過一萬像素，病理醫師標註訓練用資料十分耗時。如何能充分利用無標註的病理影像資料，減少病理醫師標註訓練資料的負擔，也是 AI 在病理影像分析中十分重要的研究議題。在本文接下來的章節，將分別介紹結合多倍率病理影像，提升病理 AI 模型的準確度，以及利用半監督式學習，減少訓練病理 AI 模型需要使用的病理標註數量。本文相關研究成果發表於⁽¹⁻³⁾。

二、結合不同倍率病理影像提升病理 AI 模型準確率

如同前一章節提到的，病理醫師進行診斷時，會來回切換不同倍率檢視病理影像，以大視野範圍的低倍率觀察病理組織的關聯與差異，以及以小視野範圍的高倍率確認細胞的紋理與排列，同時考量不同倍率下觀察到的病理特徵，能更精確地辨識腫瘤區域。因此病理 AI 模型若能有效從不同倍率與視野範圍的病理影像中，擷取更完整的病理特徵資訊，應能提升模型準確率。

由於全幅數位病理影像 (whole slide images, WSI) 尺寸巨大，受限於記憶體空間，無法以病理 AI 模型直接分析整張 WSI，需要切成 patch 分別處理後再整合分析結果。為了讓模型能夠取得細胞層級與組織層級的更完整病理特徵資訊，常見作法是在切 patch 時，以同一個中心位置，同時擷取具有局部細胞細節資訊的高倍率病理影像 patch X^c ，以及大範圍組織關聯資訊的低倍率病理影像 patch X^l ，以兩張相同位置不同倍率與視野範圍的 patch 作為模型的影像輸入。

為了能有效擷取高倍率病理影像中擷取細胞紋理資訊與低倍率病理影像中擷取組織關聯資訊，並整合完整的病理特徵提高偵測準確率。我們提出了多倍率影像異質模型，如圖 2 所示。模型分為兩個不同架構的子模型 MSUnetFusion 與 MSUSegFormer，每個子模型各有一條 Local Encoder E^L 與 Global Encoder E^G ，分別擷取高倍率病理影像 patch X_i^c 的細胞層級病理特徵與低倍率病理影像 patch X_i^t 的組織層級病理特徵。

子模型的 Local Encoder E^L 在 X^c 上擷取模型多階層不同 layer 的細胞層級病理特徵 F^c ：

$$F^c = \{F_i^c\}_{i=1}^n = E^L(X^c) \quad (1)$$

其中 i 表示 Local Encoder 的不同層 layer。同樣地，子模型的 Global Encoder E^G 在 X^t 上擷取模型多階層不同 layer 的組織層級病理特徵 F^t ：

$$F^t = \{F_j^t\}_{j=1}^m = E^G(X^t) \quad (2)$$

其中 j 表示 Global Encoder 的不同層 layer。

融合兩個不同倍率影像的特徵，便可以讓病理 AI 模型在進行影像分割時能同時考慮高倍率影像的細胞紋理特徵與低倍率影像的組織關聯特徵。然而兩張 patch 的影像倍率不同，每個像素對應到原 WSI 的座標也不相同，Local Encoder 的高倍率影像特徵圖與 Global Encoder 的低倍率影像特徵圖在 patch 坐標系上相同位置像素的特徵，表示的是原 WSI 坐標系上不同位置像素的特徵，不能直接相加或串接。高倍率影像 patch 在原 WSI 上對應到低倍率影像 patch 的中間區域，因此，我們可以使用特徵中央融合模組 (feature center fusion)，取得 Global Encoder 低倍率影像特徵圖對應高倍率影像位置的中央區域，與 Local Encoder 最後一層 layer 的特徵 F_n^t 融合得到融合特徵 $\hat{F}_n^c$ ，如圖 2(c) 所示。

首先 Global Encoder 每一層 layer 的特徵 F_j^t ，會先使用中央特徵裁切模組 (feature center cop, FCC($\cdot$))，裁切對應高倍率影像範圍的中央區域。假設高倍率影像與低倍率影像的倍率比例為 s (例如：高倍率影像為 10 倍放大倍率，低倍率影像為 2.5 倍放大倍率，則 $s = 4$)， F_j^t 特徵圖的尺寸為 $h \times w$ ，則 F_j^t 特徵圖中央中心點 $(h/2, w/2)$ ，尺寸為 $h/s \times w/s$ 的區域，對應到高倍率影像的範圍，以 FCC($\cdot$) 裁切此範圍的特徵圖，得到 $\hat{F}_j^{t\text{-cf}}$ ：

$$\hat{F}_j^{t\text{-cf}} = \text{FCC}(\text{MLP}(F_j^t)) \forall j \quad (3)$$

接著將 Global Encoder 不同層 layer 中央區域的特徵圖 $\hat{F}_j^{t\text{-cf}}$ 上採樣縮放至與 Local Encoder 最後一層 layer 特徵圖 F_n^t 相同大小後串接得到 $\hat{F}^{t\text{-cf}}$ ：

$$\hat{F}^{t\text{-cf}} = \text{Concat}\left(\left\{\text{Upsample}(F_j^t)\right\}_{j=1}^m\right) \quad (4)$$

將 $\hat{F}^{t\text{-cf}}$ 與 F_n^t 融合可得到同時具有組織層級與細胞層級病理特徵的融合特徵 F_n^c ：

$$\hat{F}_n^c = \text{MLP}\left(\text{Concat}(F_n^t, \hat{F}^{t\text{-cf}})\right) \quad (5)$$

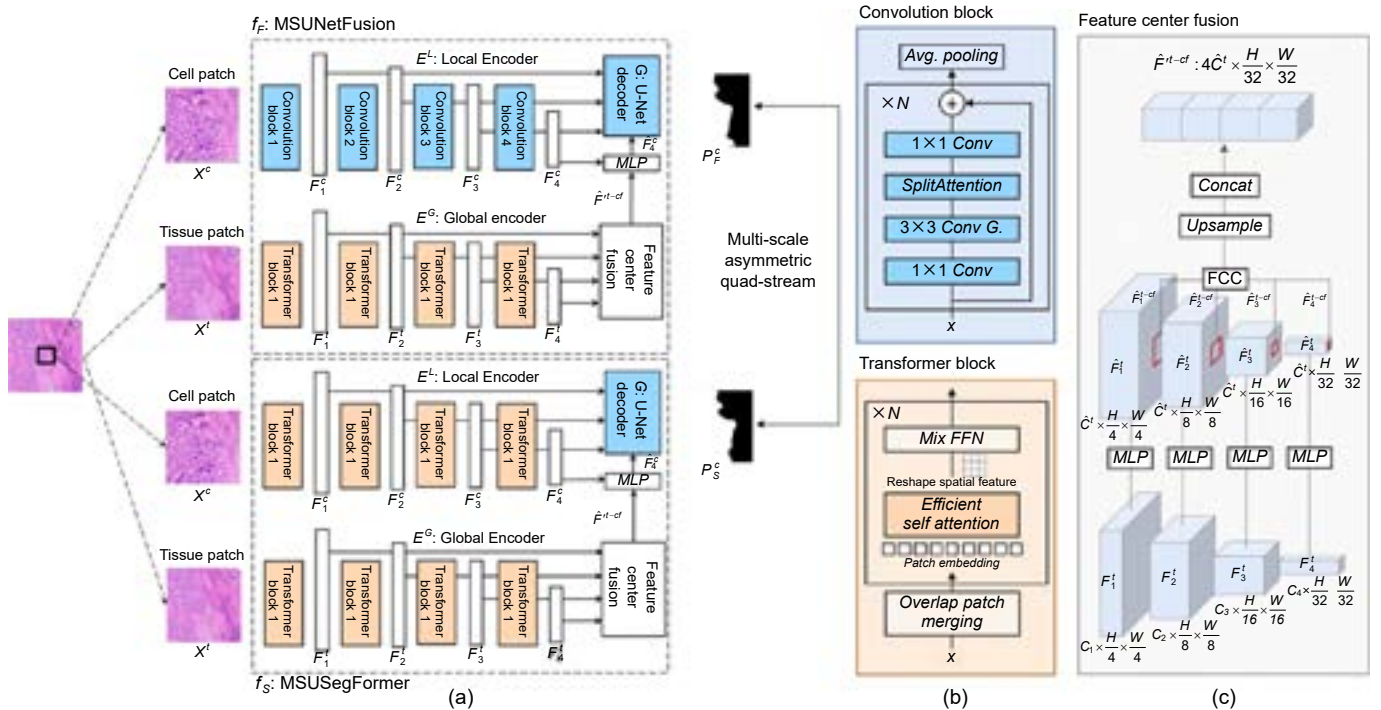


圖 2. 結合多倍率異質性模型交互監督的半監督式數位病理影像分割方法示意圖。

最後每個子模型的 U-Net-based Decoder G 整合 Local Encoder 不同層 layer 的特徵 F_i^t 與融合特徵 $\hat{F}_n^c$ ，產生腫瘤區域的影像分割 P^c ：

$$P^c = G\left(\text{Concat}(\{F_i^t\}_{i=1}^{n-1}, \hat{F}_n^c)\right) \quad (6)$$

除此之外，不同模型架構的具有不同能力，例如 CNN 架構模型擅長學習局部區域的紋理，而 Transformer 架構的全局注意力機制則能夠擷取全局的關聯。考量前述處理低倍率影像的 Global Encoder E^G 需要在中央區域擷取到周圍區域的全局組織關聯，因此 Global Encoder E^G 採用 Transformer 架構模型，而 Local Encoder E^L 則分別使用 CNN 架構模型與 Transformer 架構模型，分別組成子模型 MSUNetFusion 與 MSUSegFormer，讓兩個子模型能夠從不同視角分析病理影像。

三、利用半監督式學習減少訓練所需病理標註資料量

為了減少病理醫師準備訓練模型用的病理標註所需的時間，常見作法是利用半監督式學習⁽⁴⁻⁵⁾進行訓練。半監督式學習的訓練資料集包含少量的有標註資料集 D^l 與數量較多的無標註資料集 D^u ，通常會基於模型預測結果產生無標註資料的偽標註，再使用偽標註來計算無標註病理影像的損失函數，利用數量較多的無標註病理影像增加訓練資料量，減少訓練所需的標註資料量。

令 P_F^c 與 P_S^c 分別為本文介紹的 MSUNetFusion 與 MSUSegFormer 兩個子模型的腫瘤分割預測遮罩 (也就是分別以 MSUNetFusion 與 MSUSegFormer 使用 Eq. 6 取得的 P^c)，有標註的

病理影像以監督式學習計算損失函數 L_{SUP} ：

$$L_{\text{SUP}} = \frac{1}{|D^l|} \sum_{X^c, X^l, Y \in D^l} (l_{\text{CE}}(\mathbf{P}_F^c, Y^c) + l_{\text{CE}}(\mathbf{P}_S^c, Y^c)) \quad (7)$$

其中 Y^c 是高倍率病理影像 patch 對應範圍的影像分割標註。

而對於無標註病理影像，使用兩個子模型的分割預測遮罩產生偽標註 $\hat{Y}_F^c$ 與 $\hat{Y}_S^c$ 如下：

$$\hat{Y}_F^c = \arg \max(\mathbf{P}_F^c) \quad (8)$$

$$\hat{Y}_S^c = \arg \max(\mathbf{P}_S^c) \quad (9)$$

傳統半監督式學習常碰到的問題是偽標籤可能出錯，模型使用自己產生的錯誤偽標籤訓練，會學習錯誤的預測而影響模型準確率。為了解決這個問題，我們基於交互監督⁽⁴⁾ 提出了異質性模型交互監督的機制，讓異質性架構的兩個子模型互相使用對方產生的偽標籤計算損失函數。由於兩個子模型的架構不同，會從不同的視角分析輸入的病理影像，不容易同時發生相同錯誤，可以提升偽標註訓練無標註病理影像的可靠性。損失函數 L_{CS} 計算如下：

$$L_{\text{CS}} = \frac{1}{|D^u|} \sum_{X^c, X^l \in D^u} (l_{\text{CE}}(\mathbf{P}_F^c, \hat{Y}_S^c) + l_{\text{CE}}(\mathbf{P}_S^c, \hat{Y}_F^c)) \quad (10)$$

最後的損失函數 L 計算如下：

$$L = L_{\text{SUP}} + \lambda L_{\text{CS}} \quad (11)$$

其中 $\lambda (= 1)$ 是調整無標註影像訓練權重的超參數。藉由異質性模型進行交互監督的半監督式學習，可以有效而可靠地利用數量較多的無標註病理影像，減少訓練模型所需的病理標註資料量。

四、實驗結果

1. 實驗設置

本文中使用原發性肝癌 (liver hepatocellular carcinoma, HCC) 與大腸癌 (colorectal cancer, CRC) 的數位病理影像資料集驗證本文介紹的方法效果，並與其他半監督式學習方法進行比較。實驗經過成大醫院與高雄榮總的人體研究倫理審查委員會審查通過，並免除當事人知情同意。以回溯性收案方式蒐集成功大學醫學院附設醫院與高雄榮民總醫院去識別化的病理影像。病理切片採常規蘇木精－伊紅染色 (hematoxylin and eosin stain, H&E) 凸顯細胞與組織分布，用 Leica Aperio AT2 掃描器以 40 倍掃描成數位病理影像，並由專業病理醫師標註腫瘤區域。

肝癌資料集包含共 129 張全幅數位病理影像 (whole slide images, WSI)，隨機切分為 75 張 WSI 的訓練資料集，14 張 WSI 的驗證資料集，與 40 張 WSI 的測試資料集。大腸癌資料集共包含 65 張 WSI，隨機切分為 45 張 WSI 的訓練資料集，5 張 WSI 的驗證資料集，以及 15 張 WSI 的測試資料集。實驗以半監督式學習進行，僅使用一部分比例 WSI 的病理標註，其餘 WSI 以無標註方式用於訓練。模型準確率使用 Intersection-over-Union (IoU) 計算，採三輪固定資料切分但不同隨機種子訓練的平均與標準差。

由於 WSI 尺寸巨大，可達數萬像素，會切成多倍率影像 patch 進行訓練與測試。多倍率影像 patch 的擷取以高倍率的細胞層級影像為主，在高倍率下每隔 512 像素不重疊取一張 512×512 像素尺寸的 patch，並在低倍率下以相同中心點取一張 512×512 像素尺寸的 patch。測試時直接將每張 patch 的預測結果作為原 WSI 對應區域的預測結果。肝癌資料集細胞層級與組織層級的多倍率影像分別縮放到 10 倍與 1.25 倍的倍率，總計約有 175,000 張 patch，大腸癌資料集細胞層級與組織層級的多倍率影像則分別縮放到 10 倍與 2.5 倍的倍率，總計約有 63,000 張 patch。

其他方法的實驗分為兩個部分，一個部分是使用既有的半監督式學習方法，訓練單倍率影像模型，另一個部分是使用既有的半監督式學習方法訓練本文說明的多倍率影像模型。單倍率影像模型使用前述多倍率影像 patch 的高倍率 patch 部分進行訓練與測試。除了要進行比較的半監督式學習方法與模型不同之外，所有實驗均採用相同設定，使用帶動量的 SGD 優化器，初始 learning rate 設為 0.01，動量 (momentum) 設為 0.9，採用多項式學習率衰減策略 (Poly Learning Rate Policy, $(1-(\text{iter}/\text{maxiter}))^{0.9}$)。Batch size 設為 18，肝癌資料集訓練 15 epoch，大腸癌資料集訓練 20 epoch。偽標籤均使用 hard label，本文介紹的方法與 CPS⁽⁴⁾ 均未使用 confidence threshold 與 warm-up，UniMatch⁽⁵⁾ 使用 confidence threshold 0.95 過濾偽標籤，且前 1,000 個 batch iteration 作為 warm-up 不使用偽標籤。

2. 腫瘤區域分割結果量化比較

本文介紹的方法與其他半監督式學習方法的腫瘤區域語意分割結果，模型準確率的量化比較如表 1 與表 2 所示。表 1 呈現肝癌資料集的語意分割結果比較，本文方法比較僅使用有標註 WSI 訓練的監督式學習 (supervised)，以及兩個半監督式學習方法 CPS⁽⁴⁾ 與 UniMatch⁽⁵⁾，實驗分別使用不同比例數量的 WSI 標註。首先如表 1 第三列所示，由於 Supervised 方法僅使用少量有標註的 WSI 進行訓練，模型準確率較低，且隨著使用的 WSI 數量減少，IoU 會快速下降，如第四到七欄所示。而半監督式學習方法利用模型產生無標註影像的偽標註，可以利用更多影像訓練模型，增加模型學習到的資料，因此相較於 Supervised 方法能大幅提升準確率，如表 1 第四列與第五列所示。而我們提出的方法，利用 CNN 與 Transformer 的異質性模型架構處理不同倍率的病理影像，並且進行交互監督，讓模型能擷取病理影像中更完整的病理特徵，因此相較於其他半監督式學習方法，能更進一步提升模型準確率，如表 1 第八列所示。此外，由於本文方法的多倍率異質模型交互監督，能更全面擷取病理特徵，因此僅需少量標註資料 (3 張 WSI)，即可達到接近 Supervised 使用全部 75 張 WSI 標註的準確率。

表 1. 肝癌病理影像腫瘤區域語意分割結果比較。

		Label Data Ratio	1/25 (3 WSIs)	1/15 (5 WSIs)	1/8 (10 WSIs)	1/4 (20 WSIs)	1 75 WSIs
S	Method	Model	IoU	IoU	IoU	IoU	IoU
S	Supervised	Unet	67.8±1.2	74.0±0.3	78.7±0.8	82.2±1.1	86.3±0.5
	CPS ⁽⁴⁾		75.3±3.0	78.8±1.0	83.0±0.8	85.2±0.2	
	UniMatch ⁽⁵⁾		82.0±0.8	80.8±0.2	83.7±0.6	84.3±1.8	
M	Supervised	MSUnet	77.2±1.5	80.5±0.9	84.1±0.9	86.5±0.7	88.3±0.2
	UniMatch ⁽⁵⁾		88.1±0.9	87.9±0.3	88.1±0.8	88.2±2.5	
	MSCS (Ours)		87.7±1.4	88.5±0.4	89.3±0.4	89.7±0.5	
S	Supervised	Seg-Former	73.1±0.6	77.6±1.1	80.5±0.2	84.2±0.5	87.7±0.3
	CPS ⁽⁴⁾		76.8±1.2	81.3±0.5	81.8±0.5	85.0±0.4	
	UniMatch ⁽⁵⁾		82.1±0.5	81.9±0.9	83.2±1.4	86.5±0.5	
M	Supervised	MS-Useg-Former	77.5±1.0	80.9±1.2	85.2±0.2	87.0±0.6	88.7±0.4
	UniMatch ⁽⁵⁾		85.3±2.1	85.6±1.7	87.3±2.3	88.8±0.1	
	MSCS (Ours)		87.6±1.2	88.7±0.1	89.0±0.6	89.8±0.4	

表 2. 大腸癌病理影像腫瘤區域語意分割結果比較。

		Label Data ratio	1/22 (2 WSIs)	1/9 (5 WSIs)	1/4 (10 WSIs)
S	Method	Model	IoU	IoU	IoU
S	Supervised	Unet	60.4±0.9	68.9±1.5	72.3±0.9
	CPS ⁽⁴⁾		69.1±0.5	74.3±1.6	75.7±0.3
	UniMatch ⁽⁵⁾		72.0±0.6	73.0±1.6	75.3±0.6
M	Supervised	MSUnet	60.1±3.6	71.5±0.4	75.3±0.2
	UniMatch ⁽⁵⁾		76.9±0.9	78.9±0.2	78.0±1.6
	MSCS (Ours)		77.5±0.3	78.9±0.3	79.4±0.2
S	Supervised	Seg-Former	56.4±1.9	67.3±4.1	74.6±0.4
	CPS ⁽⁴⁾		57.9±1.2	75.1±0.3	76.4±0.3
	UniMatch ⁽⁵⁾		67.9±4.0	75.9±0.9	77.3±0.3
M	Supervised	MS-Useg-Former	61.4±3.3	72.3±1.0	75.7±0.4
	UniMatch ⁽⁵⁾		75.4±5.3	78.8±0.2	79.0±0.2
	MSCS (Ours)		77.9±0.4	79.0±0.4	79.7±0.1

同時我們並比較使用單倍率影像與多倍率影像的差異，在第一欄以 S 表示單倍率影像，M 表示使用多倍率影像。如表 1 第六列所示，Supervised 方法利用本文介紹的多倍率影像異質模型，結合高倍率影像的細胞紋理特徵與低倍率影像的組織關聯特徵，能更完整表示腫瘤區域的病理特徵，因此相較於表 1 第三列僅使用單倍率影像，使用多倍率影像異質模型

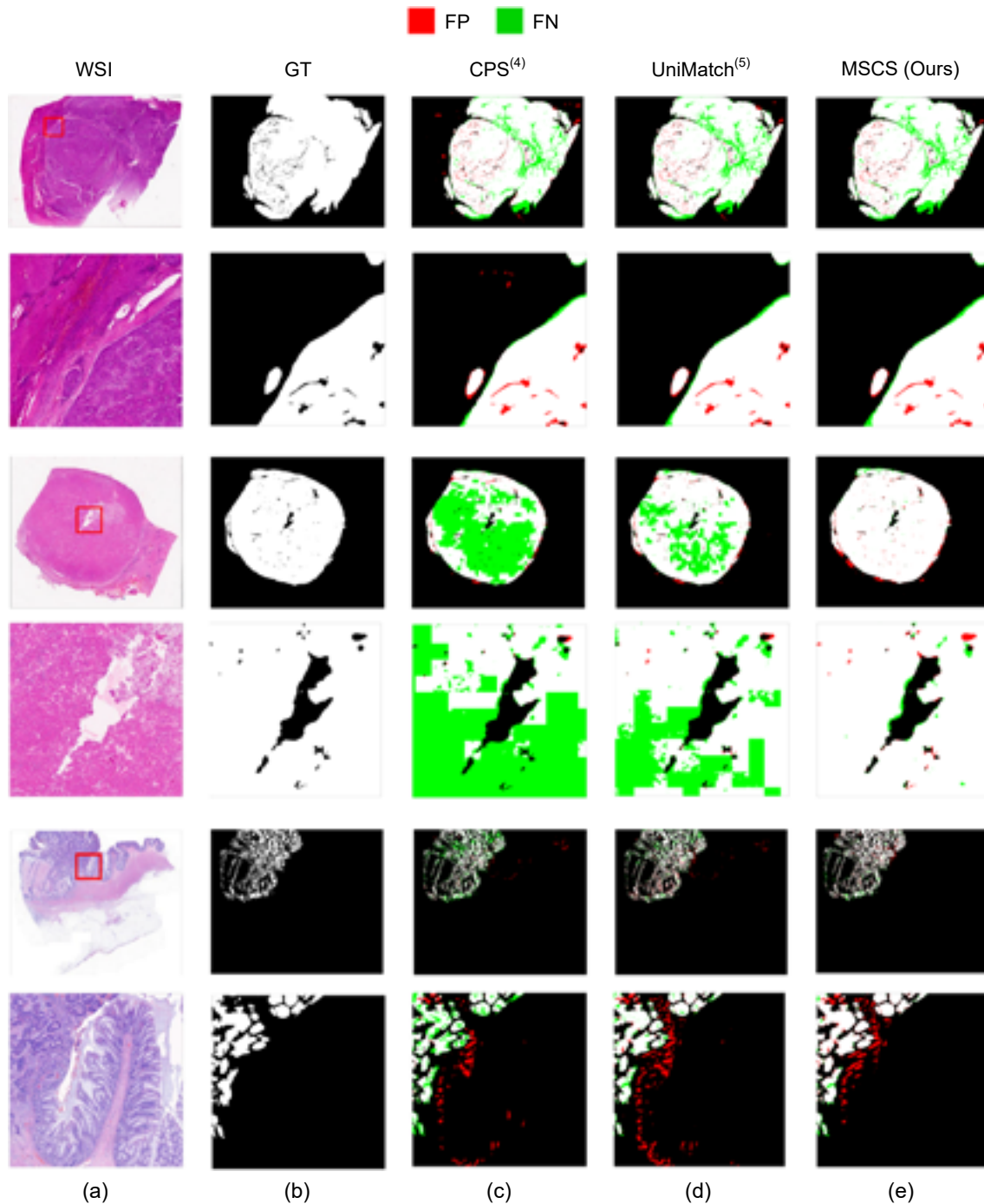


圖 3. 腫瘤區域分割結果比較。奇數列呈現全幅影像，偶數列呈現放大區域的影像，全幅影像上的紅框為放大區域。(a) 原圖。(b) GT。(c) CPS⁽⁴⁾。(d) UniMatch⁽⁵⁾。(e) 我們的方法。

的效能大幅提升。此外，由於本文方法的異質模型交互監督機制，利用兩個不同架構的模型，從不同視角產生無標註影像的偽標註，能有效降低偽標註錯誤的影響。因此如表 1 第 13 與第 14 列所示，同樣使用多倍率影像模型進行訓練，本文方法準確率仍高於其他半監督式學習方法。

大腸癌病理影像的腫瘤區域分割結果如表 2 所示。肝臟病理切片通常呈現一整塊肝臟組織，原發性肝癌的腫瘤區域常見整塊與周圍組織明顯不同的細胞密集區域。與之相對，大腸病理切片可能包含多個腺體與不規則絨毛，腫瘤區域可能分散在不同腺體區域，且有時腫瘤區域外觀與正常組織外觀差異較不明顯。因此，表 2 中所有方法大腸癌腫瘤區域的分割準確率相較於肝癌均有下降。然而本文方法利用多倍率異質模型進行交互監督，除了能更全面結合高倍率細胞細節與低倍率組織關聯的資訊，並能提升偽標註的穩定性與準確率。因此本文方法的腫瘤區域分割準確率高於監督式學習與其他半監督式學習模型的準確率，如表 2 所示。在肝臟與大腸病理影像的測試結果，顯示本文方法相較預期他半監督式學習方法，能更全面擷取病理特徵，並更有效利用無標註影像，提升模型準確率。

3. 腫瘤區域分割結果質化比較

圖 3 比較不同方法的腫瘤區域分割的視覺化結果圖。圖 3 第一列與第二列的肝癌案例，右側整塊腫瘤區域與左側正常組織區域有明顯外觀差異，腫瘤區域細胞密集而雜亂排列。因此所有方法均能完整分割出整塊肝癌腫瘤區域，僅在邊緣交界處與腫瘤區域內間質有少量誤判。而圖 3 第三列與第四列的肝癌影像，左側腫瘤區域外觀與右側正常組織區域的細胞外觀差異較小，僅使用高倍率影像較難以分辨，因此其他半監督式學習方法在腫瘤區域內部出現大量誤判為正常組織。而本文方法採用多倍率異質模型，採用能尋找全局相關性的 Transformer 架構擷取低倍率病理影像中大範圍的組織關聯特徵，搭配 CNN 學習高倍率病理影像的局部細胞紋理特徵，因此能有效分辨腫瘤區域與正常組織區域。最後一個案例是大腸癌病理影像，如圖 3 第五與第六列所示。由於大腸病理影像中有大量不規則皺褶與絨毛，細胞排列較為雜亂，與腫瘤區域差異較不明顯，因此如圖 3 第六列放大圖的左側與右下角，其他半監督式學習方法將許多正常的絨毛組織誤判為腫瘤區域。而本文提出的多倍率異質模型能取得較大範圍的組織關聯特徵，搭配異質性模型交互監督機制，提升偽標籤的可靠性與穩定性，因此模型僅在左上角腫瘤區域邊緣將少許正常組織誤判為腫瘤區域，而能成功辨識出左下角與右下角的正常組織區域。從圖 3 不同方法的腫瘤區域分割的視覺化結果圖比較，可以看見本文方法的多倍率影像異質模型與異質性模型交互監督機制，讓模型能更精確辨識腫瘤區域與正常組織區域。

五、結論

病理組織切片以顯微鏡或數位病理掃描器取得的數位病理影像，可以讓病理醫師直接觀察患者體內組織與細胞的外觀，對疾病的診斷十分重要。病理 AI 模型可以輔助病理醫師分析尺寸巨大的數位病理影像，並提供客觀的量化分析做為診斷的輔助參考資訊。為了提升病理 AI 模型偵測腫瘤區域的準確率，並減少訓練需要的病理標註資料，本文介紹結合多倍率異質性模型交互監督的半監督式數位病理影像分割方法。本方法結合 CNN 架構與 Transformer 架構的異質性模型，從高倍率與低倍率病理影像中擷取更完整病理特徵，並透過異質性模型的交互監督機制，互相產生可靠的偽標註，有效利用無標註影像資料。因此在肝癌與大腸癌病理影像資料集，本文介紹的方法均能比其他半監督式學習方法訓練的模型取得更精準的腫瘤區域分割結果。未來本文方法腫瘤區域的偵測結果，除了可以提供病理醫師作為疾病診斷的參考，也可作為後續進階分析的前處理。

參考文獻

1. Tsung-Lung Yang *et al.*, *Artificial Intelligence in Medicine*, **125**, 102244 (2022).
2. Chia Wei Chao *et al.*, *IEEE Computational Intelligence Magazine*, **18** (3), 54 (2023).
3. Min-En Hsieh, Chien-Yu Chiou, Hung-Wen Tsai, Yu-Cheng Chang, Pau-Choo Chung, *IEEE Transactions on Artificial Intelligence*, **5** (12), 6356 (2024).
4. Xiaokang Chen, Yuhui Yuan, Gang Zeng, Jingdong Wang, in Proc. IEEE/CVF Conf. *Computer Vision and Pattern Recognition*, 2613 (2021).
5. Lihe Yang, Lei Qi, Litong Feng, Wayne Zhang, Yinghuan Shi,” in Proc. IEEE/CVF Conf., *Computer Vision and Pattern Recognition*, 7236 (2023).

作者簡介

邱建毓先生為國立成功大學電機研究所博士，現為國立中正大學電機系助理教授。

Chien-Yu Chiou received his Ph.D. in Electrical Engineering from National Cheng Kung University. He is currently an Assistant Professor in the Department of Electrical Engineering at National Chung Cheng University.

詹寶珠女士為美國德州理工大學電機研究所博士，現為國立成功大學電機系特聘教授。

英文簡介

Pau-Choo Chung received her Ph.D. in Electrical Engineering from Texas Tech University, Texas, USA. She is currently a Distinguished Professor in the Department of Electrical Engineering at National Cheng Kung University.

深度材料網路：讓金屬成形模擬看得見晶粒組織

Deep Material Networks: Making Grain Structures Visible to Metal Forming Simulation

李國順、魏廷儒、陳俊杉

Guo-Chi Li, Ting-Ju Wei, Chuin-Shan Chen

金屬產品的規格書上有成分、強度與硬度，卻沒有織構，但在沖壓、軋延、抽線與鍛造的現場，凸耳、回彈與批次間的成型窗口差異，往往正是織構造成的。要讓模擬看得見織構，必須讓巨觀零件分析與微觀組織（晶粒）計算同時進行，而這種雙尺度模擬（two-scale finite element, FE²）過去因計算量過於龐大，始終停留在學術示範階段。本文介紹本團隊建立的一套以深度材料網路（deep material network, DMN）為核心的雙尺度製程模擬技術。DMN 的外形是階層網路模型，內核則是力學：每個節點都是滿足應力平衡與變形相容的均質化單元，因此只需線彈性資料即可完成訓練，換上晶體塑性等非線性材料模型後即可直接外插使用，大幅降低建模成本。在此基礎上，本團隊於計算力學旗艦期刊《Computer Methods in Applied Mechanics and Engineering, CMAME》發表〈ODMN〉文章，讓模型首度能精確描述多晶織構及其演化；並進一步發表〈TACS-GNN-ODMN〉文章，背向散射電子繞射（electron backscatter diffraction, EBSD）所量到的織構可直接推論出材料模型，換一批料不需重新訓練，對四組未參與訓練的多晶織構，應力預測的平均相對誤差低於 2%。搭配訊息傳遞介面（message passing interface, MPI）平行化的雙尺度求解器，已完成嵌入 27,676 個多晶代理模型的深抽成形模擬，以核心時間計較 FE² 等效計算加速三個數量級。模擬中沒有任何異向性擬合參數，四個凸耳的方位與高度全由各積分點的晶粒方位在成形過程中算出；模型並可直接輸出零件上任一位置的織構，供與 EBSD 量測比對，或作為下一道製程分析的輸入。同一框架已推進至全雙尺度鍛造模擬。本文說明這套技術的原理、導入所需的資料與驗證方式，以及在各類金屬製程中可切入的應用場景，並誠摯邀請國內金屬產業夥伴共同合作。

Material certificates specify composition, strength, and hardness, but not texture-yet on the stamping, rolling, wire-drawing, and forging floor, earing, springback, and coil-to-coil variation in the forming window are frequently texture-driven. Making texture visible to simulation requires solving the macroscopic part and the microscopic grain structure simultaneously, and such two-scale finite element (FE²) analyses have long been computationally prohibitive. This article introduces a two-scale process simulation technology built around the Deep Material

Network (DMN). Structured as a hierarchical network but composed of mechanically consistent homogenization units, a DMN is trained solely on linear elastic data and extrapolates directly to nonlinear constitutive laws such as crystal plasticity, sharply reducing modeling cost. Building on this, our team published ODMN paper in Computer Methods in Applied Mechanics and Engineering (CMAME), enabling faithful representation of polycrystalline texture and its evolution, followed by TACS-GNN-ODMN paper, which converts an electron backscatter diffraction (EBSD) texture measurement directly into a ready-to-use material model without retraining, with a mean stress-prediction error below 2% on four previously unseen polycrystalline microstructures. Coupled with a message passing interface parallel two-scale solver, a fully two-scale deep drawing simulation embedding 27,676 polycrystal surrogates has been completed with a core-time speed-up of three orders of magnitude over an FE² equivalent. The model contains no fitted anisotropy parameters. The position and height of the four ears follow entirely from the grain orientations at each integration point, and the texture anywhere in the part comes out as a direct output-pole figures that can be compared against EBSD measurements or carried into the next forming operation. The same framework has been extended to fully two-scale forging. We describe the underlying principles, the data and validation required for deployment, and the process applications this technology enables.

一、前言：織構，沒有寫在規格書上的性能參數

金屬材料的規格書上，寫著成分、降伏強度、抗拉強度、延伸率與硬度。但在沖壓廠、抽線廠與鍛造廠的現場，工程師都知道有一項參數規格書沒寫，卻直接影響良率：同一支料號、同樣通過驗收的兩批料，成型窗口可能相差很多；換了供應商，模具就得重調。

原因藏在肉眼看不見的尺度。金屬是由數以億計的晶粒堆疊而成，每顆晶粒都有自己的方位。經過軋延、抽拉、鍛造之後，這些方位不再是隨機分布，而是聚集在幾個特定取向附近——這就是晶體織構 (crystallographic texture)。織構讓材料帶有方向性：同一片板材沿不同方向拉伸，降伏強度與延展性並不相同。深抽杯緣的凸耳 (earring)、卸模後的回彈 (springback)、抽線後的異向性、鍛件不同部位的性能落差，追根究柢都是織構在巨觀尺度上的表現。

織構並非量不到。電子背向散射繞射 (electron backscatter diffraction, EBSD) 與 X 光繞射 (X-ray diffraction, XRD) 都是成熟的量測技術，國內各大學與法人單位都有能量測的設備。真正的關鍵在於：**量到的織構，用不進模擬**。現今業界常用成形分析使用的唯象 (phenomenological) 降伏準則，例如：鋼板常用的 Hill48⁽¹⁾、鋁板常用的 Yld2000-2d⁽²⁾，是以數個待定參數描述材料的異向性，參數由不同方向的拉伸試片與雙軸試驗擬合而得。這套方法成熟且有效，但模型裡並沒有「織構」這個變數：換一批料就得重新取樣、重做試驗、重新擬合；成形過程中織構會隨變形持續改變，固定的降伏面也無法描述這段演化。

國際上的因應之道是整合計算材料工程 (integrated computational materials engineering, ICME)：以計算模型串聯「製程－微觀組織－性能」，把大部分試誤搬到電腦裡預先完成⁽³⁾。福特汽車的「虛擬鋁鑄件」(virtual aluminum castings) 專案證實，ICME 可將產品與製程開發時程縮短 15－25%⁽⁴⁾。而要把 ICME 真正落實到金屬成形，缺的核心拼圖，是能同時解析巨觀零件與微觀組織的雙尺度模擬。

本文介紹本團隊近年建立的一套雙尺度製程模擬技術：以深度材料網路 (deep material network, DMN) 為核心，把量測到的織構直接轉換成模擬可用的材料模型，嵌入沖壓、鍛造

等成形分析中，不僅算出凸耳與回彈，也算出零件各處的纖構如何演化。相關成果發表於計算力學旗艦期刊《Computer Methods in Applied Mechanics and Engineering, CMAME》^(5, 6)，並已完成初步工業規模的驗證。

臺灣是全球金屬加工重鎮：手工具產業約有 2,400 家業者、近 5 萬名從業人員，年產值逾新台幣 1,300 億元，穩居全球前三大出口國⁽⁷⁾；扣件產業 2024 年產值約新台幣 1,540 億元，同樣名列全球前四大出口國⁽⁸⁾。加上上游的鋼鐵、線材與銅材，以及下游的鍛造、沖壓與切削加工，這條供應鏈長期是臺灣製造業的骨幹。我們相信，這套技術正是這條供應鏈往高值化邁進時所需要的工具。

二、要算得見晶粒，難在哪裡

要讓模擬「看得見晶粒」，物理基礎早已成熟，稱為晶體塑性 (crystal plasticity)。它描述的是：金屬的塑性變形來自差排在特定滑移系上的滑動；某個滑移系會不會啟動，取決於作用其上的剪應力是否達到臨界值，而這又由晶粒方位相對於外力的方向決定；滑移在造成變形的同時也讓晶格轉動，這正是纖構演化的來源。

實務上的做法，是把數百至數千顆晶粒放進一個代表性體積元素 (representative volume element, RVE)，對每顆晶粒解晶體塑性方程，同時滿足晶粒之間的力平衡與變形相容，再把結果「均質化」成一組等效的巨觀應力與剛度。以晶體塑性有限元素法 (crystal plasticity finite element method, CPFEM) 或以快速傅立葉轉換 (fast fourier transform, FFT) 為基礎的求解器 (如開源軟體 Dusseldorf Advanced Material Simulation Kit, DAMASK⁽⁹⁾) 都能把這件事算得很準。

問題出在算力成本。單一 RVE 跑完一條非線性應力之應變歷程，動輒需要數十萬 CPU 秒。若要模擬整個零件，理論上可採 FE^2 方法⁽¹⁰⁾，其方法將巨觀有限元素模型的每個積分點內嵌一個微觀 RVE，巨觀把變形梯度下傳，微觀回傳均質化應力與切線模數，兩個尺度逐步互相回饋 (圖 1)。這在數學上完全正確，在算力上卻是災難。以百萬元素等級的沖壓模型估算，數萬個積分點乘上數百個時間增量，再乘上每個增量內的平衡疊代次數，微觀求解次數輕易突破 10^8 次。即使動用整座超級電腦，也難以在工程時程內完成。「算得準卻算不動」，正是雙尺度模擬數十年來多停留在學術示範的根本原因。

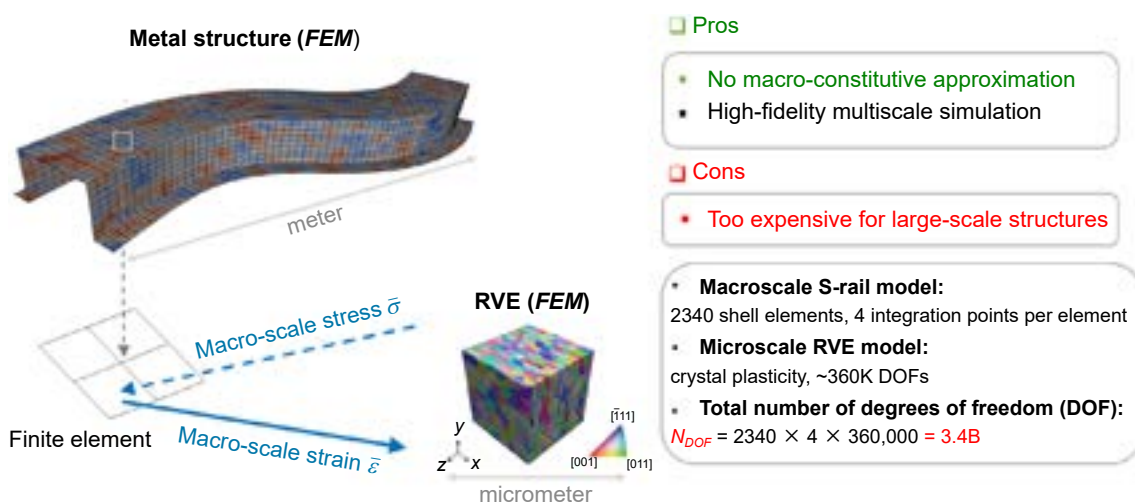


圖 1. 傳統雙尺度 (FE2) 模型說明圖。自由度太龐大 (約 34 億) 使計算無法在工程時間內完成。

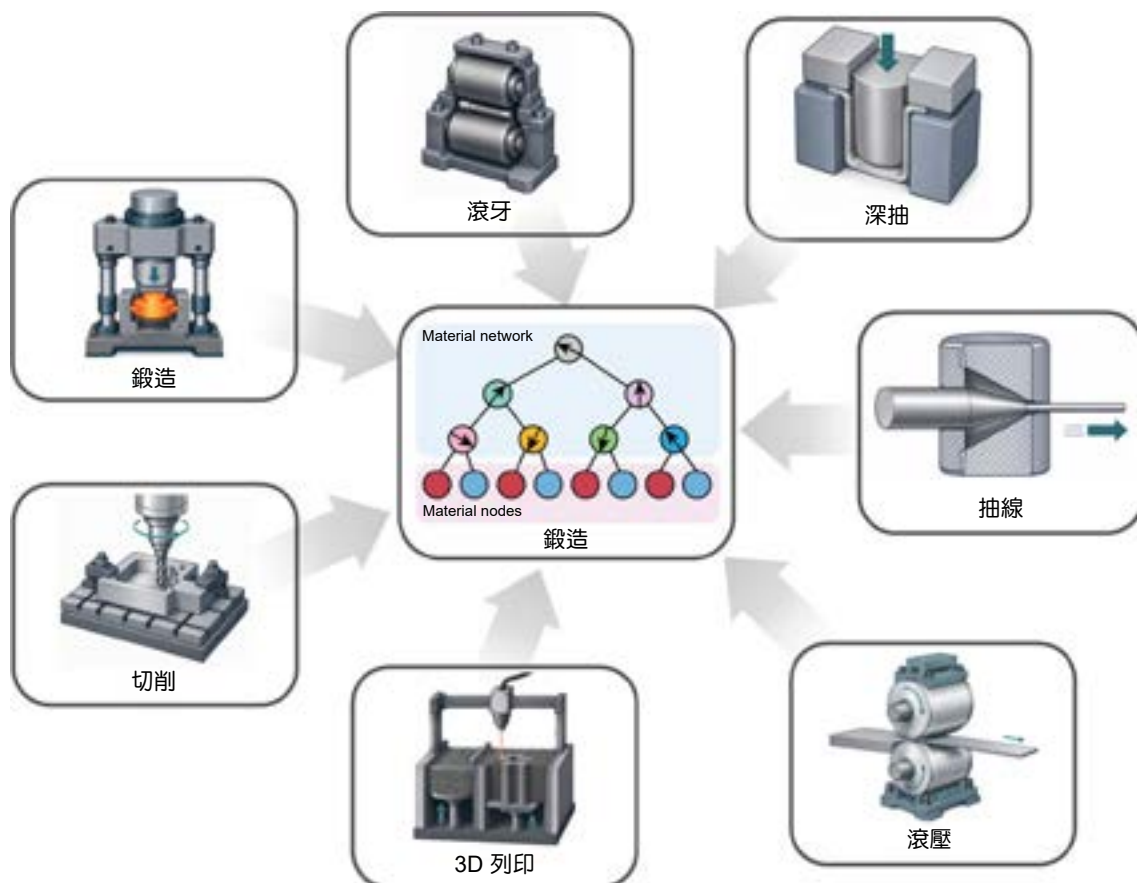


圖 2. 展示深度材料網路 (deep material network, DMN) 的通用建模概念。外圍為鍛造、滾壓、深抽、滾牙、抽線、切削與積層製造等製程，中央方框即一個 DMN，分為上下兩部分。上半部藍色區為材料網路 (material network)：它是一棵二元樹，每個節點代表一個兩相疊層的均質化單元，節點內的箭頭表示該疊層的介面法向，也就是這個單元的疊層方向，除介面法向之外，每個節點還包含體積分率。下半部粉色區為材料節點 (material nodes)：它們是樹的葉節點，每一個代表一小塊均勻材料並掛上自己的組成律，圖中以紅、藍兩色區分兩個材料相，每個葉節點上也包含體積分率。模型要學的就是這些權重與疊層方向；不同製程的差異只反映在這組參數上，因此同一套架構即可描述各類製程中材料的流動、變形與交互作用。註：外圍製程示意圖由 AI 生成，僅供參考。

要讓雙尺度模擬真正走向工程應用，關鍵在於用一個又快又準的「代理模型 (surrogate model)」取代微觀端的全域求解，而且這個代理模型必須夠快到可以被呼叫上億次，準到仍然能反映晶粒方位的物理。深度材料網路，就是為此而生 (圖 2)。

【名詞說明】RVE、均質化與切線模數

代表性體積元素 (representative volume element, RVE)：一小塊統計上足以代表整體的材料樣本；對多晶金屬而言就是一群晶粒的集合，其方位統計等於量測到的纖構。均質化 (homogenization)：把 RVE 內部不均勻的應力與應變場平均成一組等效的巨觀量，使巨觀模擬可以把這塊材料視為一個點。一致性切線模數 (consistent tangent)：應力對應變的導數，是隱式有限元素法進行平衡疊代時必須提供的量；它準不準，直接決定巨觀分析收斂得快不快、甚至收不收斂。這也是許多「AI 材料模型」無法真正嵌入商用軟體的關鍵原因。

三、深度材料網路：一個由力學組成的網路

深度材料網路是 2019 年由 Liu 與 Wu 等人提出的方法^(11, 12)。它常被歸類為「AI 材料模型」，但這個標籤容易造成誤解，它與一般用大量資料迴歸出來的神經網路，有本質上的差異。

DMN 的基本單元，是材料力學中最古老的解析解之一：兩相疊層 (two-phase laminate)。兩層材料黏合在一起承受變形時，介面上的牽引力必須連續、變形必須相容；由這兩個條件可以解析地寫出雙層結構的等效剛度。這一步不需要訓練，也不需要資料，是純粹的力學。DMN 的做法，是把這個單元堆成一棵二元樹：底層的葉節點是各個材料相，往上每一層將兩個子節點合併為一個等效介質，一路合併到樹根，就得到整個 RVE 的等效行為 (圖 3)。

模型網路要「學」的，只有每個節點的權重 (相當於體積分率) 與疊層方向。換句話說，DMN 學的是 RVE 的幾何拓撲，不是材料的反應。這一點帶來三項與傳統 AI 黑盒子神經網路截然不同的性質⁽¹³⁾：

1. **只需線彈性資料訓練**。既然學的是幾何，離線訓練只需要數百組線彈性剛度資料。在 RVE 上跑線彈性均質化即可取得，計算成本很低。訓練完成後，把節點的材料模型換成任何非線性組成律 (塑性、黏彈性、超彈性、晶體塑性)，網路就能直接外插預測，完全不需要非線性訓練資料。對導入者而言，這代表建模的前置成本從「建一個昂貴的非線性資料庫」降到「跑一批線彈性分析」。
2. **物理自洽**。DMN 由構造即滿足熱力學第二定律與能量凸性，外插到訓練範圍之外也不會產生負耗散或非物理的軟化，這是一般迴歸模型無法保證的。
3. **預測速度**。線上預測較全域模擬快三個數量級以上。

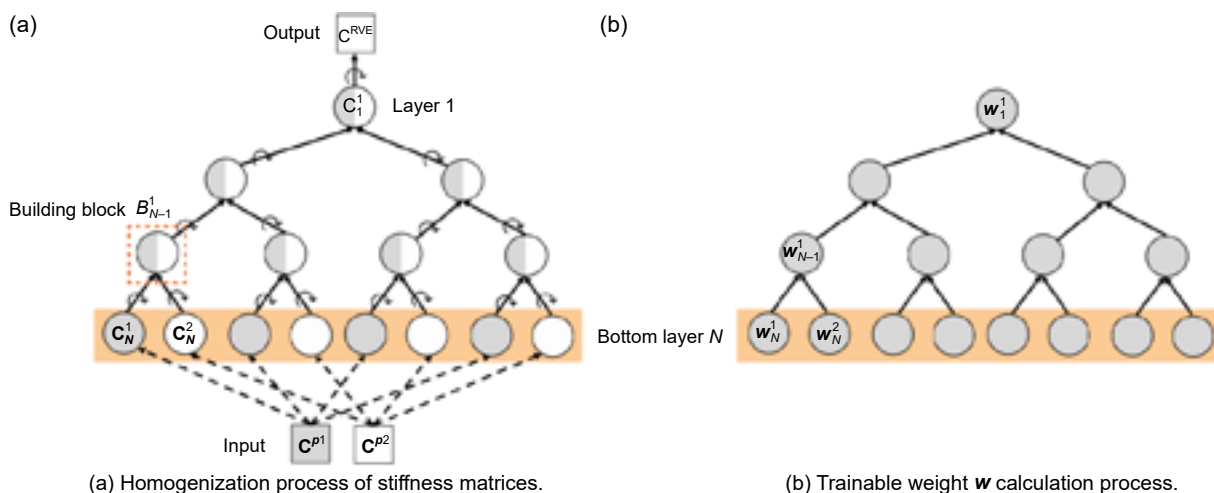


圖 3. 深度材料網路之階層式均質化架構。(a) 剛度矩陣由底層向上逐層均質化：最底層 (第 N 層) 的葉節點分別接收兩個材料相的剛度 C^{p1} 與 C^{p2} ；相鄰兩個節點構成一個兩相疊層單元 (building block)，依介面牽引力連續與變形相容的條件解析地合併成一個等效節點 C^p_l (下標 l 為層數，上標 p 為該層中的節點序號)；圖中橘色虛線框起者，即第 $N-1$ 層第一個疊層單元 B^1_{N-1} 合併後的節點。逐層向上合併至樹根即得整個 RVE 的等效剛度 C^{RVE} ；節點旁的彎曲箭頭代表該單元的疊層方向，是訓練參數之一。(b) 權重 w 的計算流程：底層各節點的權重 w^p_N 由訓練決定，上層節點的權重為其兩個子節點之和，一路加總至樹根的 w^1_1 ；權重相當於各節點所佔的體積分率 (取自參考文獻⁽¹⁴⁾)。

這些性質讓 DMN 很快走出實驗室。以隱式 UMAT¹ 在 ABAQUS² 每個積分點嵌入 DMN 的 FE-DMN 方法，已完成超過 900 萬自由度的無人機結構雙尺度分析⁽¹⁵⁾；Ansys LS-DYNA³ 團隊更與臺灣的科盛科技 (Moldex3D) 及本田技研合作，把射出成形模擬算出的纖維配向資訊轉為 DMN 代理模型，用於車用複合材料的大規模碰撞分析⁽¹⁶⁾。也就是說，DMN 在纖維複合材料領域已通過工業級驗證。下一個問題自然是：金屬材料呢？

四、ODMN：讓網路讀懂纖維

多晶金屬對 DMN 是更嚴苛的考驗。複合材料的微觀結構主要由幾何主導 (纖維形狀、配向、體積分率)；多晶金屬則不同，巨觀性能由連續分布的晶體方位主宰，而且這個分布還會隨變形持續演化。

原始 DMN 的旋轉自由度同時扮演兩種角色：既要描述晶體方位，又要決定疊層介面的方向，兩者糾纏在一起，訓練出的方位不具物理意義，自然無法描述纖維。後續的互動式材料網路 (interaction-based material network, IMN)⁽¹⁷⁾ 以 Hill-Mandel 條件重建了理論基礎、數值上更穩健，卻採「無旋轉」設計，先天無法表述晶體方位。

本團隊提出的方位感知深度材料網路 ODMN (orientation-aware interaction-based deep material network)，於 2025 年發表於計算力學領域旗艦期刊 CMAME⁽⁵⁾，補上了這塊拼圖。核心作法是「解耦」：在每個材料節點加入可訓練的晶體方位參數，專責描述纖維；另以獨立的互動機制學習 RVE 內部的應力平衡方向，專責均質化 (圖 4)。兩者分開之後，網路才真正讀得懂纖維。

僅以線彈性資料完成訓練後，ODMN 即可：

- 重建 RVE 的初始纖維，極圖與全域模擬高度吻合。
- 銜接晶體塑性組成律，精確預測單調拉伸、循環載荷與剪切下的非線性應力－應變反應，並捕捉應變速率效應。
- 追蹤大變形下的纖維演化，以及各相內部的應力分布統計。
- 相較 DAMASK 全域模擬加速二至三個數量級⁽⁵⁾。

圖 5 把 ODMN 在雙尺度模擬中一次呼叫所做的事，拆成依序執行的四個步驟：

1. 下傳 (downscaling)：巨觀有限元素在該積分點算出變形梯度 $\bar{\mathbf{F}}$ (本文中符號上加橫線者皆為巨觀量)，網路再依各節點的權重與疊層方向，把它分配成每個材料節點各自的局部變形梯度 $\mathbf{F}^i$ 。
2. 局部組成律評估：每個材料節點 $\mathbf{M}^i$ 就是一小塊均勻材料，以自己的組成律由 $\mathbf{F}^i$ 算出該節點的應力與局部切線模數；塑性部分經由塑性速度梯度 $\mathbf{L}_p^i$ 計算，彈性部分則經由右伸長張量 $\mathbf{U}^i$ ，因此可換用唯象晶體塑性、差排密度型晶體塑性或超彈性等不同模型。
3. 非線性求解：兩相疊層在介面上必須牽引力連續、變形相容，這在能量上等價於 Hill-Mandel 條件；網路以 Newton-Raphson 疊代調整節點之間的交互作用量，直到每一個疊層單元都滿足此條件為止。
4. 上傳 (upscaling)：把收斂後各節點的應力依權重平均，得到巨觀第一 Piola-Kirchhoff 應力 $\bar{\mathbf{P}}$ ；同時對其解析微分，得到圖中一併回傳的一致性切線模數 $\mathbf{L} = \frac{\partial \bar{\mathbf{P}}}{\partial \bar{\mathbf{F}}}$ ，兩者都交還給巨觀求解器，完成這一次呼叫。

1 ABAQUS 中的用戶定義材料子程式 (User-defined Material Subroutine)

2 商用有限元分析軟體

3 商用有限元分析軟體

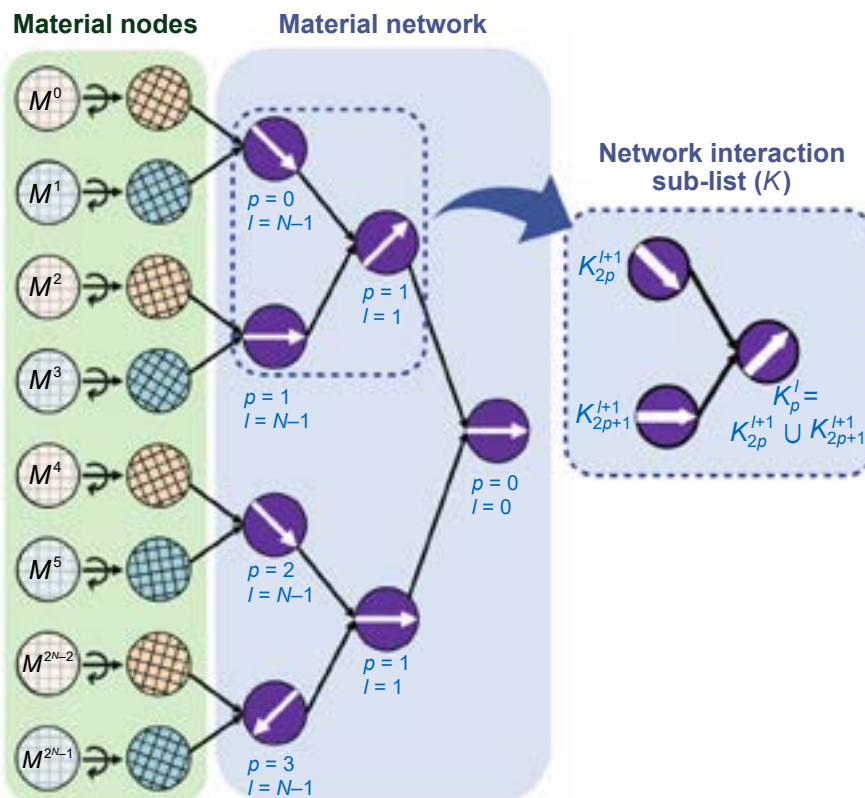


圖 4. ODMN 架構。左側綠色區為材料節點 (material nodes)，共 2^N 個，依序記為 M^0 至 M^{2^N-1} ；每個材料節點攜帶一組可訓練的晶體方位 (圖中以不同方向的網格花紋表示) 與權重，這就是「方位感知」機制，材料的纖構即由這 2^N 組方位與權重的統計來描述。中央藍色區為材料網路 (material network)：紫色圓為網路節點，圓內的白色箭頭代表該節點的交互作用方向，也就是該處的應力平衡方向；節點以 (l, p) 標定， l 為層數 ($l=0$ 為樹根， $l=N-1$ 為最底層；圖 3 的層數則自樹根第 1 層起算至最底層第 N 層，與本圖不同)， p 為該層中的序號。右側說明網路的索引關係：網路節點 (l, p) 所涵蓋的材料節點集合 K_p^l ，等於其兩個子節點集合 K_{2p}^{l+1} 與 K_{2p+1}^{l+1} 的聯集。晶體方位與交互作用方向分開訓練，正是 ODMN 能同時描述纖構與均質化的關鍵 (取自參考文獻⁽⁵⁾)。

這四個步驟在每個積分點、每個時間增量的每一次平衡疊代都要執行一次。由於切線模數是解析導出而非數值差分，巨觀平衡疊代得以維持隱式有限元素法應有的收斂速率；換句話說，ODMN 不是掛在有限元素軟體上的「AI 猜測器」，而是能像傳統材料模型一樣，透過商用有限元素軟體的正規組成律介面 (如 ABAQUS 的 UMAT) 直接接入既有分析流程的計算引擎。對於已經投資了 CAE (computer-aided engineering) 環境與工程師訓練的企業，這一點往往比任何準確度數字都重要。線上求解的數值穩定化細節，詳見本團隊投稿中的 GSH-ODMN 論文⁽¹⁸⁾。也正因為這個迴圈的呼叫次數極為龐大，代理模型每次呼叫的成本必須壓到極低，這正是第六節成本分析的主題。

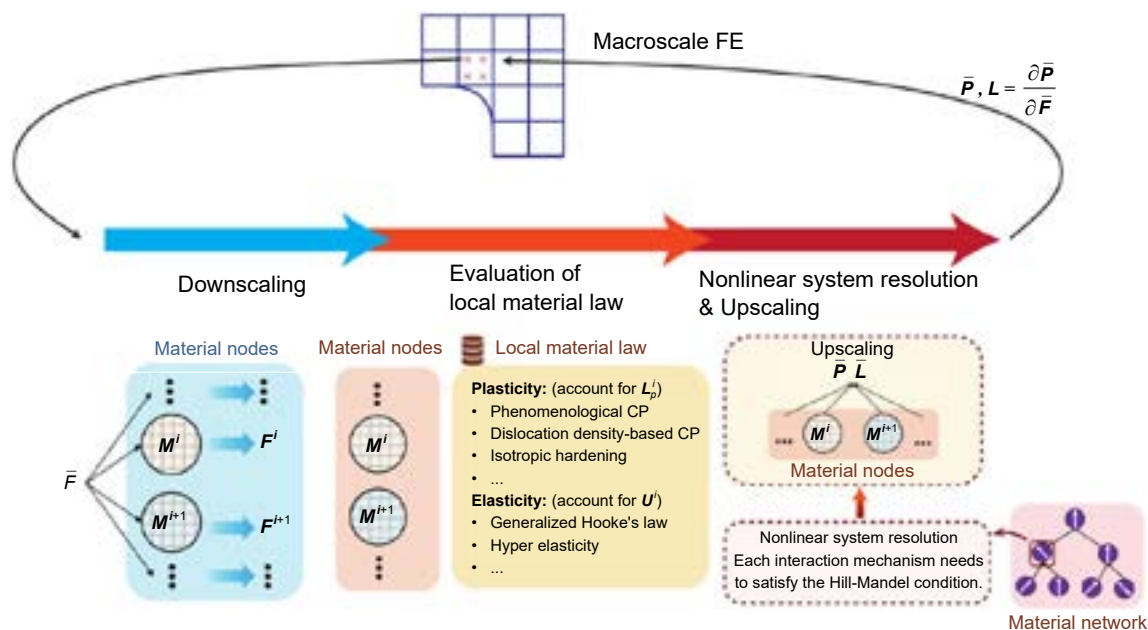


圖 5. 以深度材料網路為微觀代理之雙尺度模擬流程。巨觀有限元素模型 (Macroscale FE) 在每個積分點將變形梯度 $\bar{F}$ 下傳，網路把它分配為各材料節點的局部變形梯度 F^i ；每個材料節點 M^i 以自身的組成律更新應力，塑性部分透過塑性速度梯度 L_p^i 、彈性部分透過右伸長張量 U^i ；接著解非線性系統，調整節點間的交互作用量，使每個疊層單元都滿足 Hill-Mandel 條件；最後將各節點結果均質化上傳，回傳巨觀第一 Piola-Kirchhoff 應力 $\bar{P}$ 與一致性切線模數。圖中符號上加橫線者皆為巨觀量。此迴圈於每個積分點的每一次平衡疊代各執行一次 (取自參考文獻^(5, 14))。

五、TACS-GNN-ODMN：量完織構，模型就好了

ODMN 仍留下一個工業化的絆腳石：一個網路對應一種微觀組織。換一批料、換一種織構，就得重新訓練一次。對研究單位而言這只是多等幾個小時；對一年要驗上百批進料的鋼廠或線材廠而言，這是不可接受的。

本團隊於 2026 年再度在 CMAME 發表 TACS-GNN-ODMN 參數化代理框架⁽⁶⁾，把「換微結構組織就重新訓練」變成「量完織構直接推論」，分兩步完成 (圖 6)：

1. 織構自適應分群取樣 (texture-adaptive clustering and sampling, TACS)：EBSD 量到的可能是數萬顆晶粒的方位資料，直接建模並不經濟。TACS 先將任意的方位分布函數壓縮成一組統計等效的代表方位集，保留織構的統計特徵，但規模降到網路可處理的量級。
2. 圖神經網路推論：將代表方位集構成的晶粒交互圖輸入預先訓練好的圖神經網路 (graph neural network, GNN)⁽¹⁹⁾，直接推論出對應的 ODMN 力學平衡參數。整個過程是一次前向推論，沒有任何重新訓練。

成效如何？對四組未參與訓練的多晶織構 (材料採 AA6022-T4 鋁合金)，應力－應變曲線的平均相對誤差在循環載重下低於 2%、純剪下低於 1.5%，織構演化與局部應力統計皆與全域模擬吻合，線上計算加速 224－281 倍⁽⁶⁾。

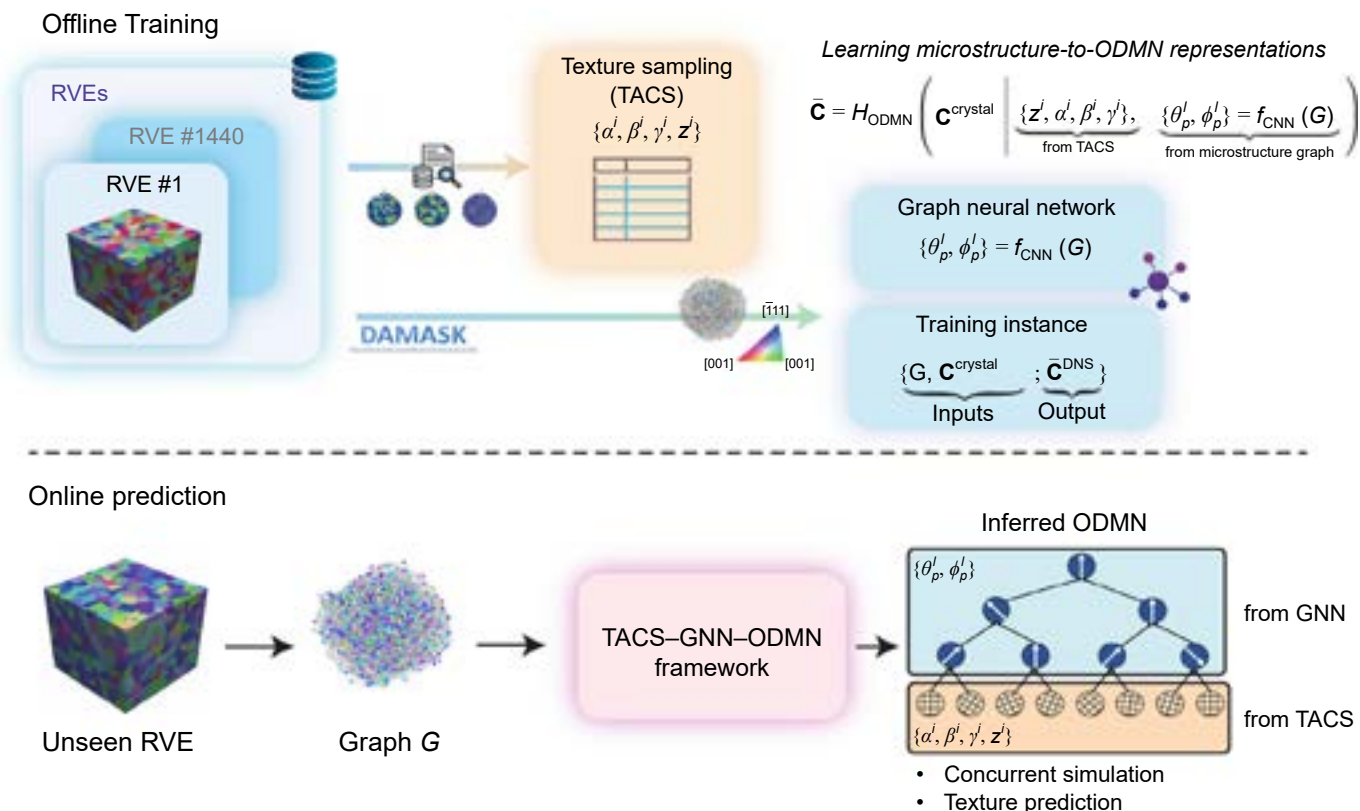


圖 6. TACS-GNN-ODMN 框架。上半部為離線訓練：由一批多晶 RVE 出發，一方面以纖維自適應分群取樣 (TACS) 把每組 RVE 的方位分布壓縮成 2^N 組代表方位與權重 α^i 、 β^i 、 γ^i 、 z^i ，其中 α 、 β 、 γ 為第 i 個材料節點的晶體方位角， z 為其權重參數；另一方面把同一個 RVE 化為晶粒交互圖 G ，輸入圖神經網路 (GNN) 推論出各網路節點的兩個方向角 θ_p^i 與 ϕ_p^i ，即圖中的 $f_{\text{GNN}}(G)$ 。兩者合起來即定義一個 ODMN：圖中的 H_{ODMN} 代表其均質化運算，把單晶剛度 $\mathbf{C}^{\text{crystal}}$ 與上述兩組參數代入網路逐層合併，算出整個 RVE 的均質化剛度 $\bar{\mathbf{C}}$ ；訓練的目標，是讓 $\bar{\mathbf{C}}$ 逼近 DAMASK 全域模擬算出的 $\bar{\mathbf{C}}^{\text{DNS}}$ 。下半部為線上預測：對一個從未見過的 RVE，只需建圖並做一次前向推論，即可得到該批材料專屬的 ODMN，直接用於併行模擬與纖維預測，不需重新訓練 (取自參考文獻⁽⁶⁾)。

這一步讓量測儀器與模擬引擎第一次真正閉環。一批鋼捲或線材進廠，切樣、拋光、量 EBSD，資料送進 TACS-GNN-ODMN，產出的就是這一批料專屬的高擬真材料代理模型，可以視為一張「數位材料卡」。下游廠商要沖什麼零件、模擬要跑什麼路徑，直接呼叫這張卡即可。進料驗收也因此從「量硬度、拉幾根試片、確認規格沒有超出」，升級為「為這批料建立一個可計算的模型」。

這裡順帶回答兩個常見的疑問。其一，EBSD 是二維截面，晶粒卻是三維的：TACS 壓縮的對象是方位分布函數 (orientation distribution function, ODF)，也就是各方位所佔體積分率的統計，而非個別晶粒的三維幾何，在統計均勻的假設下，二維截面的面積分率即等於體積分率；晶粒的三維尺寸與形貌則另循一途，依照 EBSD 統計的晶粒尺寸與長寬比建立三維合成多晶 RVE，ODMN 的互動方向即由此 RVE 訓練而得，晶粒尺寸對強度的影響則反映在以拉伸試驗校正的滑移系臨界剪應力中。其二，同一零件不同位置的纖維並不相同：雙

尺度架構的每個積分點都可指定各自的 ODMN，因此板厚方向有纖維梯度的板材可分層取樣、分層指派，鍛件與擠型材亦可分區處理。

六、規模化：MPI 平行雙尺度求解器與成形實績

有了快速且準確的微觀代理模型，還需要一套能把它掛上巨觀分析、並在數百個核心上穩定運行的求解器。本團隊開發的雙尺度求解器以 MPI (message passing interface) 平行化，可在數百個計算核心上有效擴展。

本團隊於 2026 年 7 月在德國慕尼黑舉行的第 17 屆世界計算力學大會 (WCCM-ECCOMAS 2026) 發表金屬成形的最新成果⁽²⁰⁾：深抽成形的全雙尺度模擬，巨觀模型的 27,676 個積分點，每一個都嵌入一個多晶 ODMN 代理模型，全程完成 $5 \times 10^7 - 10^8$ 次微觀求解⁴，成功重現由纖維主導的凸耳與回彈預測，以核心時間計較 FE² 架構等效計算加速三個數量級 (估算基準見表 1)。值得注意的是「每個積分點一個多晶模型」這句話的含意：這個模型裡沒有 Hill48 或 Yld2000-2d 等異向性參數。凸耳的高度與方位，是從 27,676 個積分點各自的晶粒方位一路算出來的，沒有經過任何擬合。

表 1 是這個倍數的計量基準：所有倍數皆以「核心時間」(CPU 時間 × 使用核心數) 計，在同一材料點、同一組 11 條測試路徑 (七個方向的單軸拉伸與四種預應變的反向拉壓) 下，比較全域 RVE 求解器與 ODMN 的實測成本。

表 1. 材料點層級的核心時間比較 (同一材料點、同一組 11 條測試路徑)。

求解器與設定	設備	每條路徑平均核心時間	加速倍數
DAMASK-FFT 全域 RVE (64 ³ 體素、2,000 顆晶粒)	Intel Core i9-13900K，8 執行緒	1.67 × 10 核心秒 (34.7 分鐘 × 8 執行緒)	1 (基準)
ODMN，N = 4 (16 個材料節點)	同前工作站，單核心	2.8 核心秒	約 6,000 倍
ODMN，N = 6 (64 個材料節點，深抽案例所用)	同前工作站，單核心	4.5 核心秒	約 3,700 倍
ODMN，N = 8 (256 個材料節點)	同前工作站，單核心	16.9 核心秒	約 990 倍

深抽模擬全程在國家高速網路與計算中心的單一計算節點⁵ (112 核心) 耗時 59.1 小時，合計 6,619 核心小時；若把每個積分點的 ODMN 換成全域 RVE，依表 1 的單點成本推算需要約 7.4×10^5 核心小時⁶，約 ODMN 架構的 110 倍⁷。傳統 FE²⁽¹⁰⁾ 以晶體塑性有限元素法求解微觀問題，在相同解析度下遠較 FFT 耗時⁽²¹⁾，故相對於傳統 FE² 的倍數可再高一至二個數量級。

4 視接觸與摩擦設定而異。

5 測試於國網中心 Forerunner 1 的單一計算節點 (2 顆 Intel Xeon Platinum 8480+，共 112 核心)。

6 依單點量測的每增量成本 (183,264 核心秒 ÷ 3,470 個增量，約 53 核心秒) 乘上實際發生的 5.05×10^7 次微觀求解，估計需約 7.4×10^5 核心小時。

7 以上估計皆以目前最快的全域求解器 (FFT) 為基準。整體倍數低於表一的單點倍數，是因為微觀成本降低之後，時間已由巨觀端的接觸與線性求解主導。

回到這個深抽案例本身。圖 7 是它的模型設定：胚料為半徑 85 mm、厚 1 mm 的 AA6022-T4 鋁板，沖頭半徑 50 mm、行程 55 mm，下模開口半徑 51.055 mm，胚料與沖頭的直徑比 (拉伸比) 為 1.7。巨觀模型取四分之一對稱，板厚方向分四層元素，前述的 27,676 個積分點即分布在這個網格上，每一個都掛著一個 $N = 6$ 、共 64 個材料節點的 ODMN (表 1 中深抽案例所用的設定)。分析分為下沖 55 mm、卸載與回彈三個步驟；沖頭－胚料、壓料板－胚料、胚料－下模三組接觸對與摩擦，仍由巨觀求解器以既有方式處理。

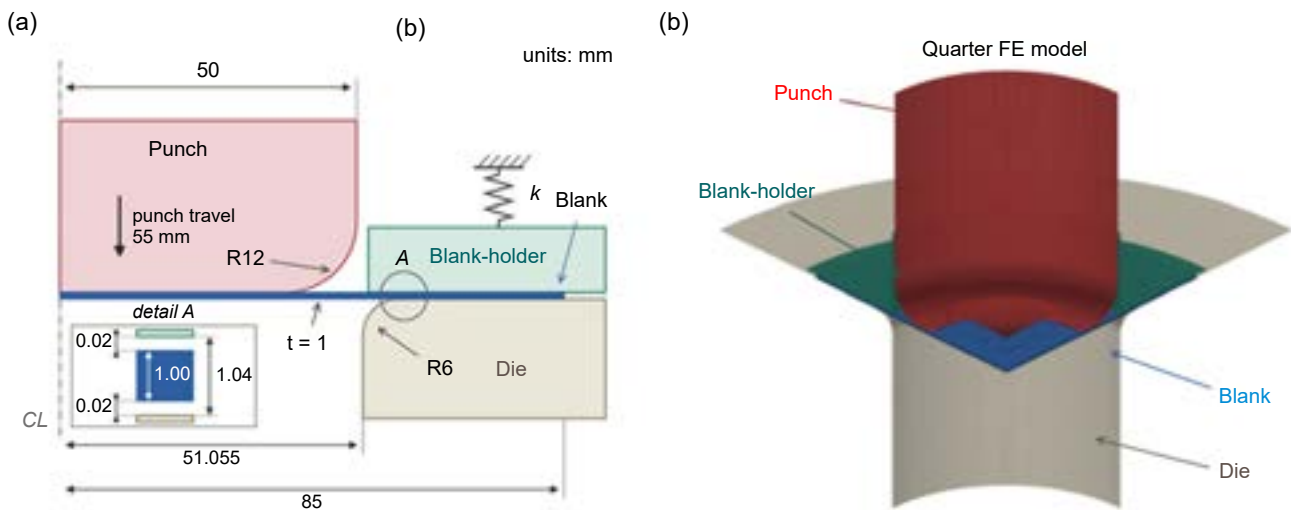


圖 7. 雙尺度深抽成形模擬的模型設定。(a) 通過對稱軸的剖面與主要尺寸 (單位: mm, CL 為對稱軸): 沖頭半徑 50 mm、鼻端圓角 R12, 下模開口半徑 51.055 mm、肩圓角 R6, 胚料半徑 85 mm、厚度 $t = 1$ mm, 沖頭行程 55 mm; 壓料板由勁度 k 的彈簧加壓, 細部 A 為壓料區的初始配置: 胚料厚 1.00 mm, 上下各留 0.02 mm 間隙, 即壓料板與下模之間 1.04 mm 的壓料間隙。(b) 四分之一對稱的三維有限元素模型, 圖中為沖頭 (紅)、壓料板 (綠)、胚料 (藍) 與下模 (灰); 胚料以 27,676 個減積分六面體元素離散 (每個元素一個積分點), 板厚方向分四層、每層 0.25 mm。

圖 8 是完整的成形歷程。沖頭下行時，法蘭區 (flange area) 的材料被逐步拉入模口，杯壁沿環向長出四處高應變區；到了最後一張，沖頭走完 55 mm 全行程、卸載回彈之後，杯緣已清楚呈現四個凸耳。四個凸耳出現的位置、長多高，完全由 27,676 個積分點各自的晶粒方位在成形過程中的轉動決定。

不過，雙尺度模擬真正獨有的產出不是杯子的形狀，而是圖 9 這一類的資訊：同一個杯件上八個位置的晶體方位分布。三種位置走上三條不同的路：杯緣的 (a)-(c) 轉動最劇烈，而且三個位置的方位各自走向不同；沖頭圓角上的 (d)-(f) 是整團方位一起轉，纖構的強弱幾乎沒變；杯底中心的 (g)、(h) 則連轉都幾乎沒轉，因為它從頭到尾貼在沖頭端面上，不再經歷什麼變形。(d)-(f) 同時給出板厚的上、下層，可以看出連同一個位置的內外表面都已經不同。這些極圖都是模擬直接輸出的結果，可以與 EBSD 的量測比對，也可以作為下一道製程分析的輸入。

同一框架已推進至鍛造製程 (見圖 10)。初步的全雙尺度鍛造模擬以 256 個 MPI 程序平行執行，單一負載歷程完成約 3.7×10^8 次微觀晶體塑性求解，其中首次嘗試即收斂的比例

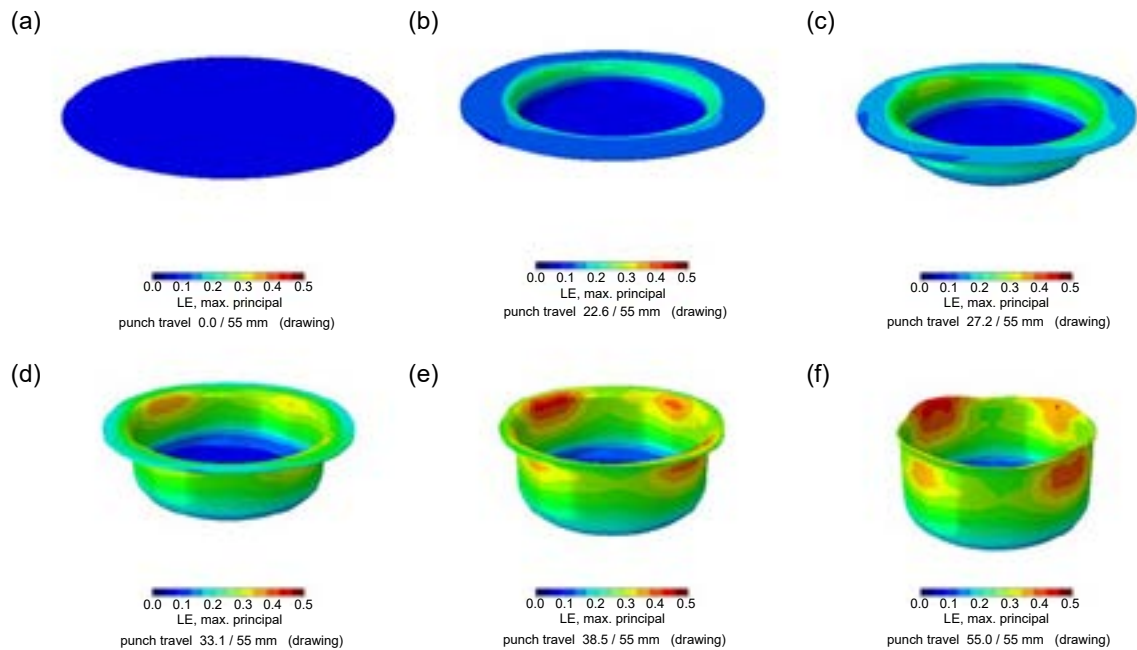


圖 8. 雙尺度深抽成形模擬的變形歷程⁸：(a) 初始胚料，沖頭行程 0 mm；(b) 22.6 mm；(c) 27.2 mm；(d) 33.1 mm；(e) 38.5 mm；(f) 沖頭走完 55 mm 全行程並卸載回彈後的最終杯件⁹。圖例為最大主對數應變 (LE, max. principal)，六張子圖共用 0 至 0.5 的同一色階。隨著行程增加，法蘭區的材料逐步被拉入模口，杯壁沿環向出現四處高應變區，杯緣則在行程後段發展出四個凸耳。

達 99.999% 以上。未於首次收斂者，程式會在該積分點內自動細化測試路徑重新求解，因此條件更嚴苛的深抽案例 (三組接觸對、法蘭起皺與回彈) 雖然首次收斂率為 99.51%，最終收斂率仍是 100%。對工業規模的非線性問題而言，這個收斂率比速度更能說明技術的成熟度。它代表求解器不會在數億次微觀計算中的某一次失敗，進而拖垮整個巨觀分析。

至於商用軟體的介接，ODMN 在巨觀求解器眼中只是一個逐點評估的材料模型：接收變形梯度與時間增量，回傳應力與一致性切線模數，介面規格與內建材料模型完全相同；幾何複雜度、接觸與摩擦仍由巨觀求解器以既有方式處理，不會因材料模型換成 ODMN 而多出新的收斂困難，前述的深抽案例本身就是這類工業級接觸問題。

鍛造件上的纖構差異更為劇烈。圖 11 展示的是鍛造件內五個代表位置的 (011) 極圖，隨壓縮量從 0.1% 到 22.1% 的演化。模具邊緣、剪切帶與死區的纖構發展路徑截然不同。同一個零件，不同位置的材料在製程結束時，實質上已經是「不同的材料」。這正是成品異向性、後續加工性與疲勞性能落差的來源，而傳統巨觀模擬完全無法提供這類資訊。

歸納而言，「ODMN 為核、GNN 泛化、MPI 擴展」的組合，已跨過從論文到應用的門檻：訓練只需線彈性資料、線上求解具二次收斂的數值品質、可接入商用求解器介面、對新纖構免重訓練。這套技術已經是「應用就緒」(application-ready) 的狀態。

8 影片連結：<https://youtu.be/kimHFQ8WMcs>

9 影片連結：<https://youtu.be/5nDLQbbIMh4>

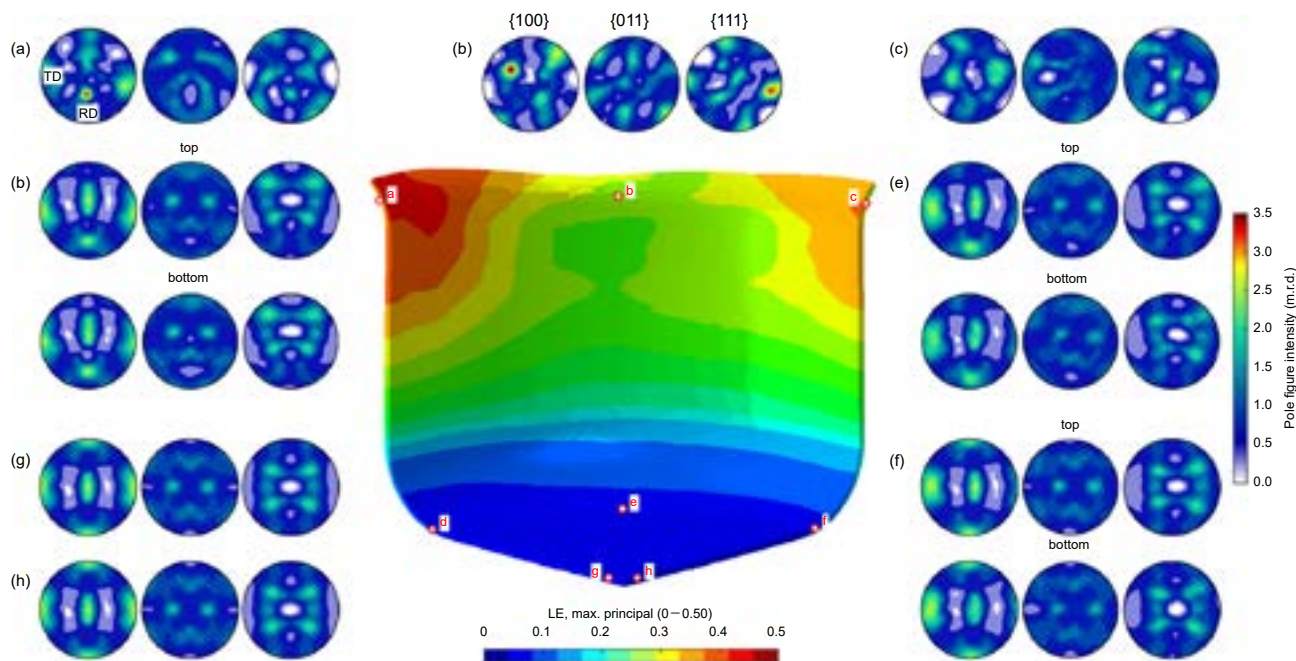


圖 9. 深抽杯件內不同位置的織構預測¹⁰ (沖頭行程 55 mm、卸載回彈後)。中央為杯件的最大主對數應變分布 (LE, max. principal, 色階 0 至 0.5), 並以 a-h 標出八個取樣位置: (a)-(c) 在杯緣, (d)-(f) 在沖頭圓角 (杯底轉角) 處; (g)、(h) 則在杯底中心附近。周圍每一列是該位置 ODMN 所繪製的 $\{100\}$ 、 $\{011\}$ 、 $\{111\}$ 極圖; 其中 (d)、(e)、(f) 分別給出板厚最上層 (top, 沖頭側) 與最下層 (bottom, 下模側) 兩組, 以顯示厚度方向的差異, 其餘位置各取一層。所有極圖共用右側同一色階, 單位為 m.r.d. (multiples of a random distribution, 數值愈高代表該方位愈集中); (a) 的第一張極圖標示了 RD (縱軸) 與 TD (橫軸), 其餘各圖方位相同。

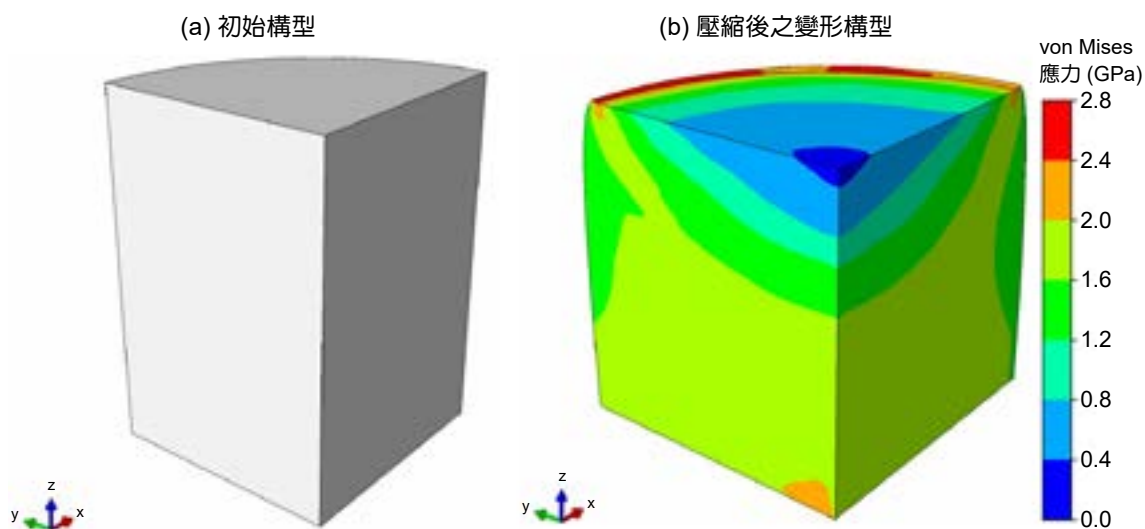


圖 10. 全雙尺度鍛造模擬初步結果 (1/4 對稱模型): (a) 初始構型; (b) 壓縮後之變形構型與內部 von Mises 應力分布。模擬以 256 個 MPI 程序平行執行, 共完成約 3.7×10^8 次微觀晶體塑性求解。

¹⁰ 影片連結: <https://youtu.be/tlmsYDwFILw>

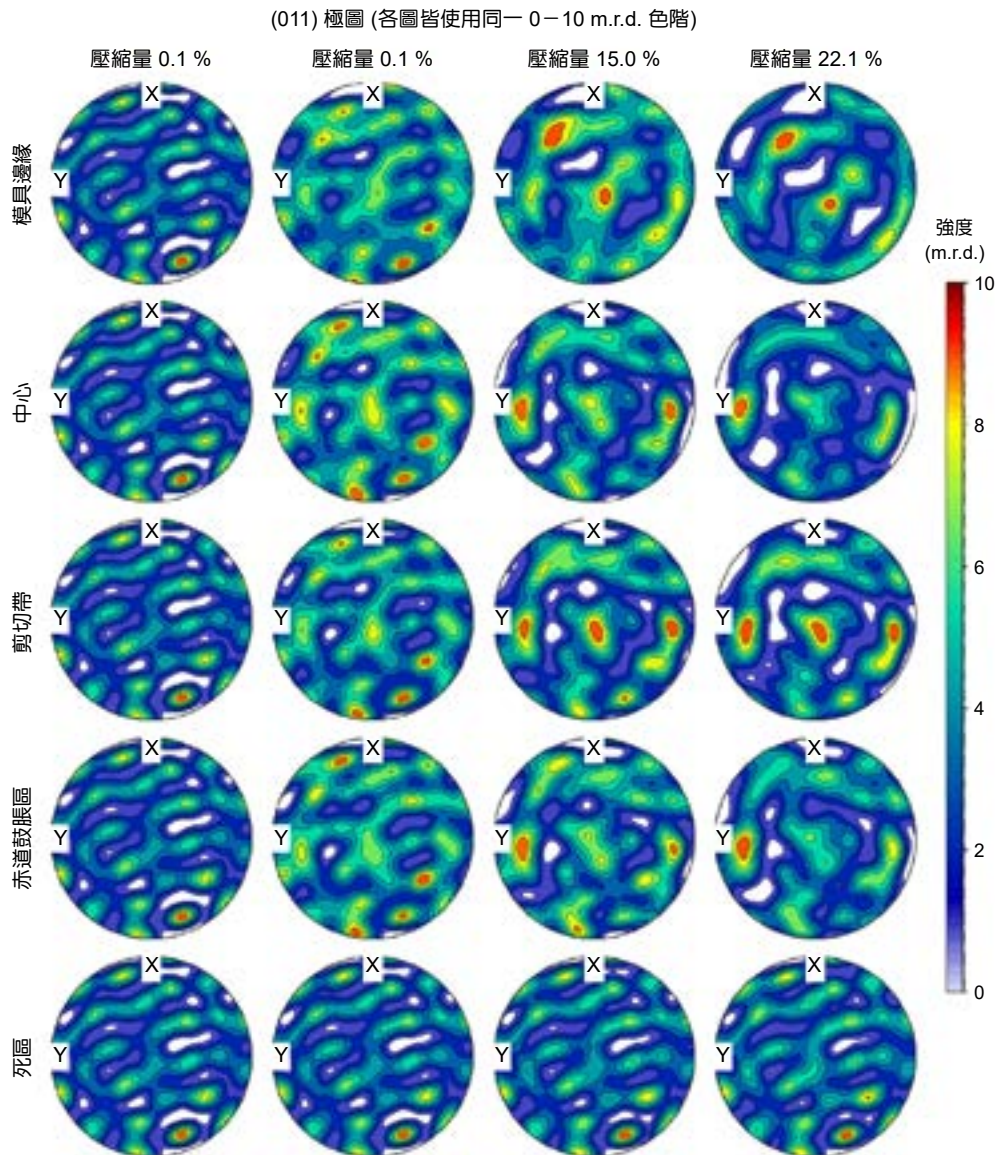


圖 11. 鍛造件內五個代表位置 (模具邊緣、中心、剪切帶、赤道鼓脹區、死區) 之 (011) 極圖隨壓縮量 (0.1 % — 22.1 %) 之演化。所有極圖共用右側同一色階，單位為 m.r.d. (multiples of a random distribution，數值愈高代表該方位愈集中)。圖中 X、Y 為試件座標軸，X 朝上、Y 朝左。各位置纖構發展路徑迥異，此類資訊唯有雙尺度模擬能夠提供。

七、導入需要什麼：資料與驗證

企業評估一項新技術時，最實際的兩個問題是：企業要提供什麼資料？憑什麼相信結果？

第一個問題的答案，是這套技術與現行做法最大的差別所在：它的輸入是組織量測，不是曲線擬合。企業提供的是「這批料的晶粒長什麼樣、方位怎麼分布」，模型自己算出降伏面該長什麼樣，而且是隨變形一路改變的降伏面。多數企業手上原本就有的品保資料，如：拉伸試驗、硬度、金相等，都能直接派上用場。

第二個問題的答案，在於把「校正」與「驗證」分清楚。校正就是用實驗資料決定模型參數，驗證則是用模型去預測沒有參與校正的實驗，再與實際結果比對。兩者若混在一起，就會得到看起來很準、實際上什麼都預測不了的模型。本團隊的作法是：滑移系的臨界剪應力與硬化參數，只用最單純、最便宜的單軸拉伸試驗校正 (因為這些是材料的本質性質，與試件形狀無關)；校正完成後即凍結參數，直接拿去預測深抽的凸耳高度、鍛造件的鼓脹形狀與各處組織，這些預測沒有任何一個參數是為它們調整過的。表 2 整理了各項資料在框架中的角色。

表 2. 導入所需資料與驗證項目。

資料／量測項目	角色	在框架中的用途
EBSD 方位圖、XRD 宏觀組織 (極圖/ODF)	建模輸入	建立 RVE 方位統計；TACS 壓縮為代表方位集
金相、晶粒尺寸與形貌	建模輸入	建立三維合成 RVE 的晶粒尺寸與形貌統計；RVE 的線彈性均質化資料決定模型的拓撲參數
材料成分與製程履歷 (軋延／退火／抽線條件)	建模輸入	判定組織來源與適用範圍
單軸拉伸曲線 (含不同方向、 r 值)	參數校正	校正滑移系臨界剪應力與硬化參數
反向拉壓、循環載荷試驗	驗證	包辛格效應與背應力，與回彈預測直接相關
液壓鼓脹、成形極限圖 (FLD)	驗證	雙軸與各種應變路徑下的成形極限
成形後各位置的 EBSD	驗證 (關鍵)	組織演化預測，唯有雙尺度模擬能比對
殘留應力量測 (XRD)	驗證	殘留應力場與回彈
成品幾何量測 (3D 掃描、凸耳高度、回彈角)	驗證	巨觀成形結果
既有試模與量產紀錄 (含不良品)	驗證	模型適用範圍與誤差界定

值得一提的是最後兩類資料的價值。企業累積多年的試模紀錄與不良品資料，過去只能作為經驗傳承；在這套框架中，它們是驗證模型最有力的素材，尤其不良品，往往比良品更有資訊量。而模擬的回饋也是雙向的：一個沖壓件切幾個位置做 EBSD 是可行的，切一百個位置不可行；模擬提供的是全場的組織與應力演化，量測提供的是幾個位置的錨點。用少數幾個位置的量測驗證模型，再用模型去看整個零件，這才是「製程虛擬學生」真正的意義。

八、可切入的製程場景

表 3 整理了本技術可切入的製程、對應的微觀組織課題，以及虛擬學生可以回答、而傳統巨觀模擬回答不了的問題。

表 3. 雙尺度虛擬孿生於各金屬製程的應用場景。

製程	現場常見課題	背後的微觀組織機制	虛擬孿生可回答的問題
深抽／沖壓	凸耳、回彈、局部頸縮、批次差異	軋延織構造成的塑性異向性與其演化	凸耳高度與方位、回彈補償量、成形極限、胚料形狀最佳化；換一批織構不同的料，成型窗口要如何調整
熱軋／冷軋 (板材、箔材)	道次規劃、下游成形性不穩定	織構隨道次與退火演化	織構控制策略、道次分配對下游成形性的影響、異向性指標預測
滾軋成形／ 螺紋滾造	牙底填充不足、疲勞強度分散	表層劇烈剪切造成的織構梯度與加工硬化	預成形 (胚料尺寸) 設計、表層織構與殘留應力分布、疲勞性能的組織依據
抽線／盤元 加工	扭轉試驗開裂、中心破裂、後續冷鍛開裂	抽線纖維織構與中心區三軸拉應力	模具半角與減面率的組合、中間退火時機、線材異向性對後續冷鍛成形性的影響
冷鍛／熱鍛	充填不良、模具負載、成品異向性	鍛流線與晶粒方位演化	預成形設計、模具負載預測、成品各部位的異向性與後續加工性
熱處理／退火	再結晶織構不易掌握	變形儲存能分布與再結晶織構	變形織構與儲存能的空間分布，作為再結晶模擬的輸入
切削加工	表面完整性、疲勞壽命分散	表層晶粒細化與殘留應力	切削參數對表層組織與殘留應力的影響
積層製造	成品異向性、後處理規劃	凝固織構、柱狀晶成長	異向性評估、後處理與熱處理製程設計

三個具體情境，或許更能說明它的價值。

情境一：讓進料差異說得清楚

一家沖壓廠長期受困於「A 廠的料好沖、B 廠的料不好沖，但兩家的成分與拉伸規格都落在同一個範圍」。用傳統試驗說不清楚，因為差別在織構，而織構不在規格書上。透過 TACS-GNN-ODMN，兩批料各量一次織構、各建一張數位材料卡、各跑一次成形模擬，差異即可量化呈現，也能據此與供應商討論具體的織構指標，而不是含糊的「品質不穩」。

情境二：把試模次數換成模擬次數

一個新的深抽件或鍛造件從首次試模到量產，反覆修模是常態。若能在開模前先用雙尺度模擬掃過胚料形狀、壓邊力與潤滑條件的組合，把凸耳與回彈的趨勢先算出來，省下的不只是模具費，更是交期。以 ICME 的國際經驗換算 (開發時程縮短 15—25%⁽⁴⁾)，效益可直接反映在成本與上市時程上。

情境三：從「控成分」到「控織構」

對上游鋼廠與線材業者而言，這套技術最長遠的意義，是讓織構從一個學術名詞變成可設計、可承諾的產品規格。當您能算出「這樣的織構配上客戶的製程，成型窗口會有多寬」，材料就從商品變成解決方案，這是高值化最實在的一條路徑。

九、未來展望

在研究層面，這套框架仍有幾個正在推進的方向。其一是網路深度與成本的取舍：以廣義球諧函數 (generalized spherical harmonics, GSH) 將量測織構直接壓縮成模型材料節點的初始方位，使較淺的模型深度即可精確表述複雜織構，並同步處理熱—力耦合問題⁽¹⁸⁾。其二是再結晶與晶粒成長：這類機制不只是讓既有晶粒轉動，還會產生原本不存在的新方位、並重新分配各方位的體積分率，模型的材料節點因此必須能夠增減與改變權重。其三是更嚴苛的製程條件，例如：冷軋的大應變與板厚方向織構梯度、切削的劇烈剪切與熱軟化，以及顯式動力求解器的介接。其四是以自監督預訓練建立多晶材料的基礎模型，使新織構的推論更為泛化。

在應用層面，本團隊誠摯邀請國內金屬產業的先進共同合作，把這套技術帶進產線。為了讓合作從一開始就務實可行，以下說明典型的推進方式：

1. 界定場景：由業者提出一個具體、有痛感、且已累積歷史資料的製程課題 (例如某深抽件的凸耳與回彈、某鍛件的晶粒流向、某批線材的抽線異常)。我們通常建議第一個題目選「老問題」而非「新問題」。因為老問題已有大量試模與量產紀錄，正好用來驗證模型。
2. 建立材料數位分身：以既有的材料試驗與織構量測資料，或由本團隊協助規劃量測，建立該批材料專屬的 ODMN 代理模型。
3. 建置雙尺度虛擬孿生：完成該製程的全雙尺度模擬，與既有的試模／量產結果交叉驗證，並明確界定模型的適用範圍與誤差。
4. 製程優化與技術移轉：以虛擬孿生進行參數掃描與最佳化，並將模型、工作流程與操作方式移轉至業者端，讓貴公司的工程師能自行運用。

業者提供的資料與衍生模型均以保密協定規範，智慧財產歸屬於合作契約中議定。舉凡深抽／沖壓、軋延、滾軋成形與螺紋滾造、抽線、鍛造、熱處理、切削加工、積層製造。只要產品性能與晶粒組織、織構息息相關，都非常歡迎與我們聯繫 (聯絡人：陳俊杉特聘教授，dchen@ntu.edu.tw)。

十、結論

本文的結論可歸納為四點。第一，在方法上，深度材料網路把「快」與「準」的矛盾解開了：模型學的是 RVE 的幾何拓撲而非材料反應，因此只需數百組線彈性資料即可完成離線訓練，換上晶體塑性組成律後即可直接外插；ODMN 再以方位感知機制讓網路讀懂織構與其演化，TACS-GNN-ODMN 則讓量測到的織構經一次前向推論即產出材料模型，對四組未參與訓練的多晶織構，應力預測的平均相對誤差低於 2%，線上計算加速 224—281 倍。第二，在規模上，這套代理模型已通過工業級尺寸的驗證：深抽成形的全雙尺度模擬在 27,676 個積分點各嵌入一個多晶 ODMN，全程完成 $5 \times 10^7 - 10^8$ 次微觀求解，以核心時間計較 FE² 架構等效計算加速三個數量級；整個模型裡沒有任何異向性擬合參數，四個凸耳的方位與高度都是從各積分點的晶粒方位所算出來的；鍛造案例則以 256 個 MPI 程序完成約 3.7×10^8 次微觀晶體塑性求解。第三，在產出上，雙尺度模擬獨有的價值不只是杯子的外形與凸耳：在同一個深抽杯件上取八個位置，極圖顯示杯緣的晶粒方位轉動最劇烈；沖頭圓角處是整個方位一起轉動；杯底中心則幾乎沒有變化，連同一位置的板厚上、下層都已經不同。鍛造件五個代表位置的 (011) 極圖，演化路徑更是截然不同。這些織構都是模擬的直接

輸出，可與 EBSD 量測比對，也可作為下一道製程與性能評估的輸入，而傳統巨觀模擬完全無法提供這類資訊。第四，在數值品質上，ODMN 解析導出一致性切線模數，可經由商用軟體的標準組成律介面 (如 ABAQUS 的 UMAT) 直接接入既有分析流程；鍛造案例首次嘗試即收斂的比例達 99.999% 以上，條件更嚴苛的深抽案例為 99.51%，未收斂者全數由局部子步驟回復，最終收斂率為 100%。

從 2019 年 DMN 問世，到 ODMN 讓網路讀懂纖構、TACS-GNN-ODMN 讓量測直接產出模型，再到嵌入數萬個多晶代理模型的全雙尺度成形模擬，「以晶粒組織為本的製程虛擬孿生」已從願景走到應用門前。對金屬產業而言，這代表老師傅口中的「手感」，第一次有機會被寫成算得動的物理；也代表纖構這個過去只存在於實驗室報告中的名詞，可以成為產品規格與製程設計的一部分。臺灣擁有完整的金屬加工聚落與深厚的現場經驗，正是把這項技術轉化為產業競爭力最合適的土壤。我們期待與產業夥伴攜手，讓下一個世代的金屬製程，在電腦模擬中先成形一次。

誌謝

本研究承國家科學及技術委員會 (計畫編號 111-2221-E-002-054-MY3、112-2221-E-007-028、113-2221-E-007-005、114-2221-E-002-010-MY3) 經費支持，並感謝國家高速網路與計算中心 (NCHC) 及國震中心-台大土木人工智慧研究中心提供計算資源。

參考文獻

1. Rodney Hill, *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, **193** (1033), 281 (1948).
2. Frederic Barlat et al., *International Journal of Plasticity*, **19** (9), 1297 (2003).
3. National Research, C., *Integrated Computational Materials Engineering: A Transformational Discipline for Improved Competitiveness and National Security*. 2008, Washington, DC: The National Academies Press.
4. John Allison, Mei Li, C. Wolverton, XuMing Su, *The Journal of The Minerals, Metals & Materials Society (TMS)*, **58** (11), 28 (2006).
5. Ting-Ju Wei, Tung-Huan Su, Chuin-Shan Chen, *Computer Methods in Applied Mechanics and Engineering*, **441**, 117977 (2025).
6. Ting-Ju Wei, Tung-Huan Su, Chuin-Shan Chen, *Computer Methods in Applied Mechanics and Engineering*, **461**, 119250 (2026).
7. 李士勛,《更上一層樓篇》手工工具產業的「攻」與「守」：工業級工具,臺灣手工工具敲開國際市場,中華民國對外貿易發展協會, (2025). Please refer to the website: <https://www.taitra.org.tw/cp.aspx?n=37>
8. 廖鴻瑞, 金屬製品產業年鑑—扣件篇, 財團法人金屬工業研究發展中心 (2025).
9. Franz Roters et al., *Computational Materials Science*, **158**, 420 (2019).
10. Frederic Feyel, Jean-Louis Chaboche, *Computer Methods in Applied Mechanics and Engineering*, **183** (3-4), 309 (2000).
11. Zeliang Liu, C.T. Wu, M. Koishi, *Computer Methods in Applied Mechanics and Engineering*, **345**, 1138 (2019).
12. Zeliang Liu, C.T. Wu, *Journal of the Mechanics and Physics of Solids*, **127**, (2019).
13. Ting-Ju Wei, Wen-Ning Wan & Chuin-Shan Chen, *Multiscale Science and Engineering*, **8**, 1 (2026).
14. 魏廷儒, 深度材料網路於微結構學習與多尺度建模之研究, 國立臺灣大學土木工程學系, 國立臺灣大學: 臺北市 (2026).
15. Gajek, S., M. Schneider, and T. Bohlke, *Computer Methods in Applied Mechanics and Engineering*, **384**, 113952 (2021).
16. Haoyan Wei et al., *Journal of Engineering Mechanics*, **149** (3), 04023003 (2023).
17. Van Dung Nguyen and Ludovic Noels, *Computer Methods in Applied Mechanics and Engineering*, **389**, 114300 (2022).
18. Ting-Ju Wei and Chuin-Shan Chen, *From Compact Orientation Distributions to Texture Evolution: GSH-Initialized Orientation-Aware Deep Material Networks*. (2026).

19. Jimmy Gaspard Jean *et al.*, *Computational Mechanics*, **75**,113 (2025).
20. Chuin-Shan Chen and Ting-Ju Wei, “Two-Scale Modeling of Polycrystalline Metals with Orientation-Aware Deep Material Networks”, *17th World Congress on Computational Mechanics & 10th European Congress on Computational Methods in Applied Sciences and Engineering (WCCM-ECCOMAS 2026)*, July 19-24 (2026).
21. Arun Prakash and Ricardo Lebensohn, *Modelling and Simulation in Materials Science and Engineering*, **17** (6), 064010 (2009).

作者簡介

李國頌先生現為國立臺灣大學土木工程學系博士生。

Guo-Chi Li is currently a Ph.D. student in the Department of Civil Engineering at National Taiwan University.

魏廷儒先生為國立臺灣大學土木工程學系博士，現為國立臺灣大學土木工程學系博士後研究員。

Ting-Ju Wei received his Ph.D. in Civil Engineering from National Taiwan University and is currently a Postdoctoral Researcher in the Department of Civil Engineering at National Taiwan University.

陳俊杉先生為美國康乃爾大學土木與環境工程博士，現為國立臺灣大學土木工程學系特聘教授暨材料科學與工程學系特聘教授。

Chuin-Shan Chen received his Ph.D. in Civil and Environmental Engineering from Cornell University, USA. He is currently a Distinguished Professor in the Department of Civil Engineering and the Department of Materials Science and Engineering at National Taiwan University.

從異常偵測走向理解與行動：人工智慧於智慧製造檢測的技術演進與可信部署

From Anomaly Detection to Understanding and Action: The Evolution and Trustworthy Deployment of AI for Smart Manufacturing Inspection

林士智、賴尚宏

Shih-Chih Lin, Shang-Hong Lai

異常偵測是智慧製造品質檢測的關鍵技術，但實務仍面臨缺陷樣本稀少、未知異常難以事先定義，以及模型難以跨產品與跨感測條件泛化等挑戰。本文從人工規則、正常資料學習、合成異常、少樣本與零樣本方法出發，說明視覺基礎模型、語意－結構融合，以及統一與跨域模型的技術演進，並延伸討論視覺語言問答、代理式取證與可信部署。本文強調，任務準確度、異常定位、因果視覺依賴、誤報與漏檢，以及部署安全，應分別評估，才能建立可驗證、可追溯，並能協助工程師決策的異常理解系統。

Anomaly detection is an important technology for quality inspection in smart manufacturing, yet practical deployment remains difficult because defect samples are scarce, unknown anomalies cannot be exhaustively defined, and models often generalize poorly across products, sensors, and production conditions. This article reviews the evolution from rule-based inspection and normal-only learning to synthetic anomaly generation, few-shot and zero-shot detection, visual foundation models, semantic-structural fusion, and unified cross-domain systems. It further discusses the transition from anomaly detection to anomaly understanding through vision-language question answering, structured reasoning, and agentic evidence acquisition using zooming, normal-reference retrieval, and difference maps. Particular attention is given to trustworthy deployment. Task accuracy, anomaly localization, causal visual dependence, false alarms, fatal misses, and operational safety are related but distinct properties and should be evaluated separately. A reliable inspection system should therefore provide verifiable and traceable evidence, preserve data and model lineage, and support engineers in making risk-aware decisions rather than merely producing fluent explanations.

一、前言

在現代製造環境中，產品外觀、組裝品質與設備狀態，直接影響良率、可靠性與生產成本。常見的異常包括刮痕、裂縫、污染、凹陷、元件漏裝、位置偏移及晶圓圖案異常。這些缺陷所佔比例通常極低，卻可能在後續製程中被放大，進而造成重工、報廢、客訴，甚至安全風險。運用相機、顯微鏡、深度感測器與人工智慧建立自動光學檢測 (automatic optical inspection, AOI) 系統，已成為智慧製造的重要基礎。

異常偵測與一般影像分類有本質上的差異。分類模型可從大量標註影像中學習固定類別；工業現場卻往往只有充足的正常資料，真實缺陷不僅稀少，也無法涵蓋未來可能出現的所有未知異常。因此異常偵測的核心並非記住每一種缺陷，而是建立正常產品應有的外觀、結構與關係，再找出偏離正常狀態的證據。相關綜述將主要挑戰歸納為訓練資料稀少、感測模態多樣及異常層級複雜三項；近年的方法也逐步從針對每項產品個別訓練，轉向運用基礎模型先驗進行跨產品泛化^(1, 2)。

當異常偵測走入真實產線後，還必須進一步思考第四個問題：模型分數高，是否就代表能夠安全部署？現代異常分析不再只關注「能否找到缺陷」，還必須檢查模型是否使用正確的影像證據、能否適應新產品與資料漂移、是否會產生大量誤報，以及錯誤判斷會不會在後續診斷流程中持續放大。如圖 1 所示，技術目標正從規則式判定，逐步走向多模態理解、主動取證與可信部署。本文其餘章節安排如下：第二章界定異常分析的任務層級與實務挑戰；第三、四章依時間序回顧從人工規則到基礎模型的技術演進；第五章討論如何從偵測走向理解與主動取證；第六章聚焦可信部署的評估方法；第七章整理跨產業應用；第八、九章展望未來並作結。



圖 1. 異常偵測與理解技術的演進。本文示意圖均由作者繪製。

二、異常偵測的任務層級與實務挑戰

2.1 從判斷、定位到異常理解

工業異常分析可分為四個層次。第一是影像層級判斷，輸出產品為正常或異常的判定與分數；第二是異常定位，以熱度圖或邊界框指出可疑區域；第三是像素層級分割，精確標示缺陷的形狀與範圍；第四是異常理解，進一步辨識異常類型、外觀特徵、與正常參考的差異及可能影響。這四項能力並不會自然同時成立：模型可能正確判定整張影像為異常，卻定位到錯誤區域；也可能產生看似合理的文字說明，卻未真正使用影像證據。

2.2 結構異常與邏輯異常

結構異常通常可從局部外觀發現，例如裂縫、孔洞、污染、微小刮痕或病灶；邏輯或語意異常則涉及元件的數量、位置與關係，例如漏件、錯裝、組裝順序不合理，或「每個元件單獨看來正常，但整體關係錯誤」。前者需要細緻的紋理與幾何特徵，後者則需理解產品整體結構及元件規則。CSAD 利用基礎模型建立元件區域，再分析元件分布、局部外觀與整體關係，說明異常偵測已從「像素是否不同」逐步走向「場景是否合理」⁽³⁾。圖 2 彙整了上述任務層級，以及結構異常與邏輯異常的差異。

2.3 真實產線使問題更為複雜

研究資料集多半在固定相機、光線與背景下建立，真實產線則可能面臨批次差異、設備漂移、材料反光、姿態變化、產品改版與新型缺陷。異常區域通常面積小、對比低，正常加工痕跡與缺陷外觀也可能十分相似。另一方面，高產量產線每天處理大量產品，即使誤報率看似不高，也可能造成可觀的複檢負擔。以每日檢測十萬件的電子組裝線為例，誤報率僅百分之一即產生一千件複檢案例；每件耗時三十秒，相當於每日額外投入八小時以上人力。反之，若每千件產品中有一件缺陷品未被攔截，每日即有上百件缺陷品流入客戶端。因此，模型不僅要有良好的平均表現，也必須在指定誤報率、固定精確率、實際生產節拍及長期資料漂移下，維持穩定的檢出能力⁽²⁾。

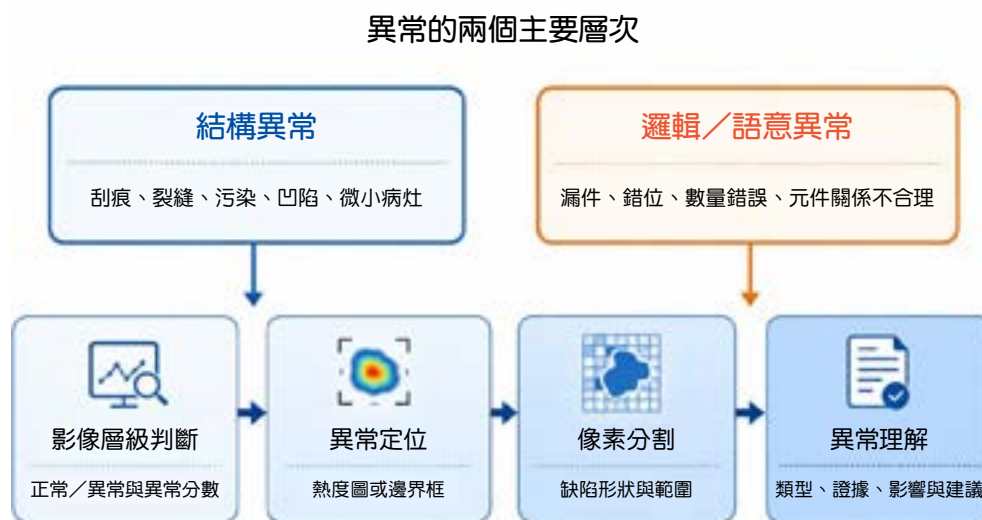


圖 2. 異常分析的任務層級，以及結構異常與邏輯異常的差異。

三、從人工規則走向正常資料深度學習

3.1 規則式機器視覺

早期檢測系統多採用灰階閾值、顏色範圍、邊緣、幾何尺寸、紋理統計與樣板比對。這類方法運算快速、容易解釋，在產品位置、光線與缺陷型態固定的任務中仍相當有效；但一旦產品、材料或環境改變，工程師就必須重新設定規則，而且也難以處理不規則、低對比或未知缺陷。

3.2 正常資料學習：重建、蒸餾與記憶庫

深度學習方法多以正常影像為主要學習資料。重建式方法先讓模型學習正常產品應有的外觀，再以輸入影像與重建結果之間的差異找出可疑區域；知識蒸餾方法則比較預訓練教師模型與只看過正常資料的學生模型之間的特徵差異；記憶庫方法會保存正常影像的局部特徵，測試時再尋找最相近的正常參考。MVTec AD 是廣泛使用的工業異常基準資料集，提供正常影像、缺陷影像與缺陷區域標註，常用於比較不同方法。PatchCore 的概念像建立一座「正常特徵資料庫」：先把正常影像切成許多局部區塊並保存其特徵，測試時若某個區塊找不到相近的正常樣本，就提高它的異常分數。DRAEM 則先在正常影像上人工製造缺陷，再訓練模型學習恢復正常外觀並同時標出異常位置⁽⁴⁻⁶⁾。這三類方法各有適用場域：重建式方法直觀，且可輸出模型所推定的正常外觀，便於檢視模型行為；蒸餾式方法毋須重建整張影像，推論速度通常較快；記憶庫方法在正常樣本充足時定位精準，惟記憶庫的容量與檢索時間會隨產品種類增加而成長。

3.3 模型容量並非越大越好

重建模型存在一項基本矛盾：解碼器過弱時，正常結構無法正確恢復，容易產生誤報；解碼器過強時，又可能連缺陷一起重建，使原本應該突出的異常差異反而消失。對晶圓或週期性紋理而言，模型既要記住正常結構的位置關係，又不能把異常細節原封不動地複製回去。作者團隊提出的 SQUAD，將連續影像特徵轉成有限的離散表示，可將其理解為限制模型僅能以一組貼近正常樣態的「基本構件」重建影像，藉此降低直接複製缺陷的機會⁽⁷⁾。圖 3 比較了兩種重建方式的實際差異。更一般而言，可透過限制模型容量、在中間特徵加入適度擾動，以及加強困難區域的學習，在「保留正常對應」與「避免複製缺陷」之間取得平衡。這也說明，異常偵測模型並不是越大越好。

3.4 合成異常補足資料不足

真實缺陷不足時，可在正常影像上合成刮痕、污染、孔洞、紋理替換或擴散生成缺陷，並同步取得像素層級遮罩。作者團隊的多合成動態加權方法整合不同生成方式，使模型能學習更多樣的異常型態⁽⁸⁾；LFQUIAD 進一步結合合成擴散修補與量化表示，處理少樣本情境⁽⁹⁾。然而，合成影像看來逼真，不表示一定有助於真實檢測；模型可能只學到生成工具留下的痕跡。因此，合成資料應同時評估外觀合理性、缺陷位置、真實測試效益，以及不同缺陷類別是否出現不對稱的誤報或漏檢代價。若模型對合成缺陷的分數遠高於真實缺陷，通常顯示其學到的是生成工具的痕跡，而非缺陷本身的特徵。

為使表 1 所示的演進幅度更加具體，可以最常用的 MVTec AD 基準之影像層級 AUROC 為例：針對單一產品類別訓練的記憶庫方法 PatchCore 約 99.1%⁽⁵⁾，重建判別式方法 DRAEM 約 98.0%⁽⁶⁾；改以單一模型同時處理十五類產品的統一模型 UniAD 約 96.5%，而完全不使用目標產品訓練資料的零樣本方法 WinCLIP 約 91.8%（兩者分別見 4.3 節與 4.1 節）。這組數字顯示，在單一類別且正常資料充足的設定下，公開基準表現已接近飽和；統一與零樣本方法則以數個百分點的差距，換取大幅降低的資料蒐集與模型維護成本。惟須強調，上述皆為公開基準上的排序指標，並不同於特定產線的誤報與漏檢表現，後者仍應依第六章的部署導向指標另行評估。

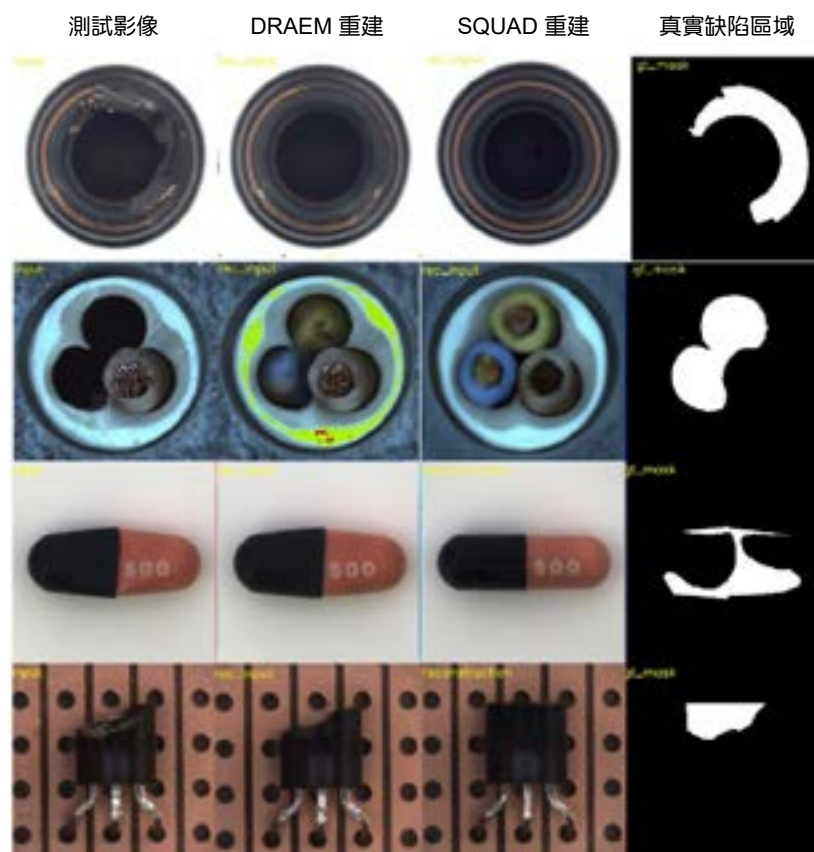


圖 3. 重建式異常偵測的實際案例。由左至右為測試影像、DRAEM 重建結果、SQUAD 重建結果與真實缺陷區域。當模型把異常一併重建時，缺陷差異可能被削弱；SQUAD 則嘗試恢復較接近正常的外觀，使異常更容易被發現。影像取自作者團隊 SQUAD 研究⁽⁷⁾，重新排版。

表 1. 視覺異常偵測技術演進與實務特性比較。

方法階段	主要資料	代表概念	優點	主要限制
規則式	少量樣本+人工規則	閾值、邊緣、樣板	快速、易解釋	產品或環境改變後需重設
正常資料學習	正常影像	重建、蒸餾、記憶庫	無須大量缺陷標註	容量與對應的權衡、逐類維護成本
合成異常	正常影像+人工缺陷	剪貼、紋理、擴散生成	可取得像素遮罩	合成痕跡偏差、類別風險不均
少／零樣本	少量或無目標資料	CLIP 語意、DINO 結構、正常參考	縮短新產品導入時間	微小缺陷感受力與跨域穩定性
統一與跨域	多類別／來源域資料	共享模型、凍結骨幹	降低模型維護成本	未知產線適應與來源域過度擬合
異常理解	影像+文字+知識	視覺語言模型 (VLM)、問答、結構化推理	可描述、定位與提出假設	幻覺、模板捷徑與證據不一致
主動取證代理	影像+參考+工具	放大、檢索、差異圖	模擬工程師查證流程	工具路徑與證據使用未必一致
可信部署	測試資料+風險成本	FA/FM、因果審計、穩定性	依操作風險選擇模型	校準、漂移與跨場域驗證

四、基礎模型驅動的下一波技術演進

大型預訓練模型改變了異常偵測長期依賴單一產品資料的方式。CLIP (Contrastive Language-Image Pre-training, 對比式語言－影像預訓練) 同時從大量影像與文字中學習，因此能把「正常」、「破損」或「缺少元件」等文字概念與影像內容對應起來；DINO 則是一類透過自監督學習 (self-supervised learning) 從大量影像中學習的視覺模型，較擅長保留形狀、邊界與局部紋理等結構資訊。簡言之，CLIP 著重於回答「這是什麼、語意是否相符」，DINO 則長於辨識「何處的結構與周圍相異」。分割基礎模型則可進一步協助切出物件與元件區域^(10, 11)。近期綜述將可供泛化的先驗整理為五類：視覺先驗、語言與推理先驗、幾何與多模態先驗、通用任務先驗，以及生成異常先驗⁽²⁾，如圖 4 所示。這些先驗可互補使用，但仍需在目標場域驗證是否真的能降低誤報，並保留對微小缺陷的感受能力。

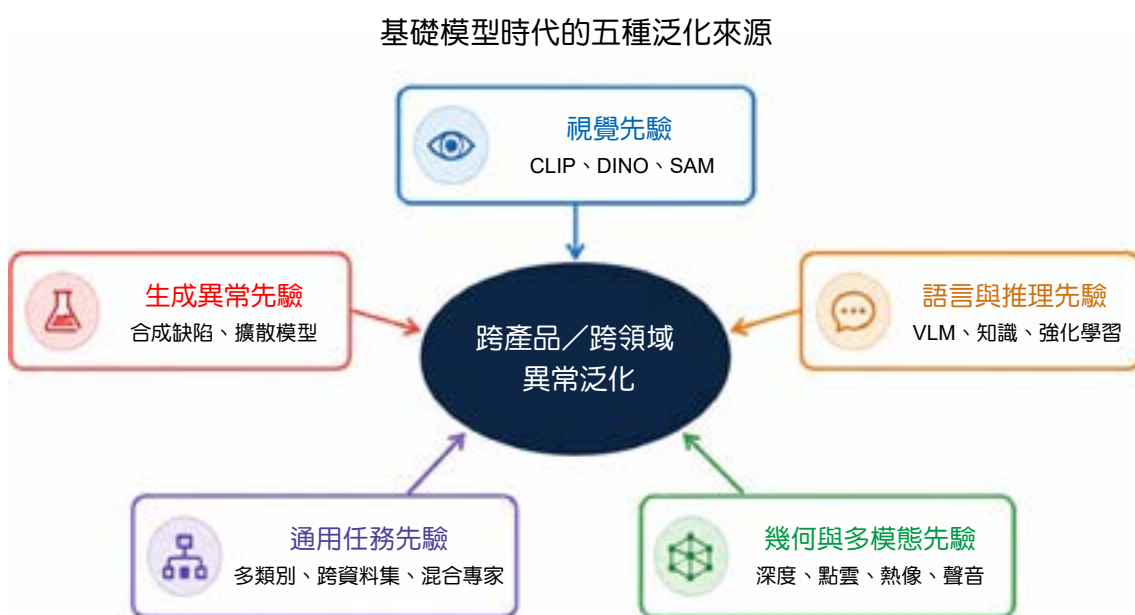


圖 4. 基礎模型時代的五種泛化來源。

4.1 少樣本與零樣本異常偵測

少樣本 (few-shot) 方法利用一張或少量正常參考影像，建立產品的正常表示；零樣本 (zero-shot) 方法則在完全沒有目標產品訓練資料時，直接借助預訓練知識進行判斷。WinCLIP 是把 CLIP 用於工業異常檢測的代表性方法：它利用「正常」與「異常」文字提示，再配合影像的局部區域特徵，判斷哪裡可能有缺陷⁽¹²⁾。AnomalyCLIP 進一步讓異常提示不必綁定特定產品類別，使同一套語意概念更容易跨產品使用⁽¹³⁾。作者團隊的 CLIP-FSQA 則將 CLIP 提供的語意資訊與離散量化表示結合，讓少量正常樣本也能形成較穩定的正常參考⁽¹⁴⁾。對產線而言，這類方法的實質意義在於縮短新產品的導入時間：過去每項新產品往往須蒐集數百張影像並重新訓練，如今僅需少量正常樣本或一段文字規格，即可先建立可用的初版模型，再隨資料累積逐步強化。

4.2 語意與結構必須互補

CLIP 擅長理解「正常」、「破損」或「缺少元件」等整體語意概念，但對很小的刮痕、低對比病灶或細微幾何變形，未必足夠敏感；DINO 類模型則較能保留局部形狀與紋理連續性。圖 5 示意兩者如何互補：由 CLIP 提供「正常／異常」的語意判斷，再由 DINO 補上局部結構證據，最後分別產生像素層級的異常圖與整張影像的異常分數。實作上，結構資訊只以小幅度加入，避免破壞 CLIP 原本的語意能力；同時把「找位置」與「判斷整張影像是否異常」分開處理，讓模型既能看清小缺陷，也不會因局部熱度圖的調整而意外改變整體判斷⁽¹⁵⁾。

語意與結構互補的跨域異常分析流程



圖 5. 以語意與結構互補進行異常定位與影像排序的概念流程。

4.3 從單一產品走向統一與跨域模型

傳統作法通常為每項產品建立一個模型；當產品類別增加後，訓練、校正與版本維護成本也會快速上升。統一異常偵測的目標，是讓「一個模型」處理多種產品。UniAD 是此方向的代表架構之一，讓不同產品共用同一套模型⁽¹⁶⁾；作者團隊的 UniLEAD 則進一步採用較輕量的共享設計，讓不同產品共用大部分能力，只保留少量必要的專門處理，降低模型數量與維護成本⁽¹⁷⁾。當廠內產品達數十項時，統一模型可將數十套模型與校正紀錄收斂為一套共享能力加上少量產品專屬設定，版本管理與異常追蹤因而簡化。更進一步的跨域偵測，要求模型在來源產品上訓練後，能直接轉移至新產品、不同成像條件，甚至跨到醫療影像。透過凍結大型基礎模型、只訓練少量介面，以及分開處理語意與結構資訊，可降低模型過度記住來源資料的問題；作者團隊的 UFMADS 與 BSMAD 也顯示，這類輕量融合方式可延伸到跨器官或跨成像模態的異常偵測與分割^(15, 18)。

圖 6 以真實影像說明跨資料集的遷移結果：上半部為模型於 VisA 訓練後直接測試於 MVTEC AD，涵蓋瓶子、纜線、膠囊、螺絲與紋理表面等工業物件；下半部則為反向設定。雖然兩個資料集的產品外觀與成像條件差異很大，模型在未曾見過的目標資料集上，仍能將較高的異常反應集中在真正有問題的區域，而非背景或正常結構⁽¹⁹⁾。

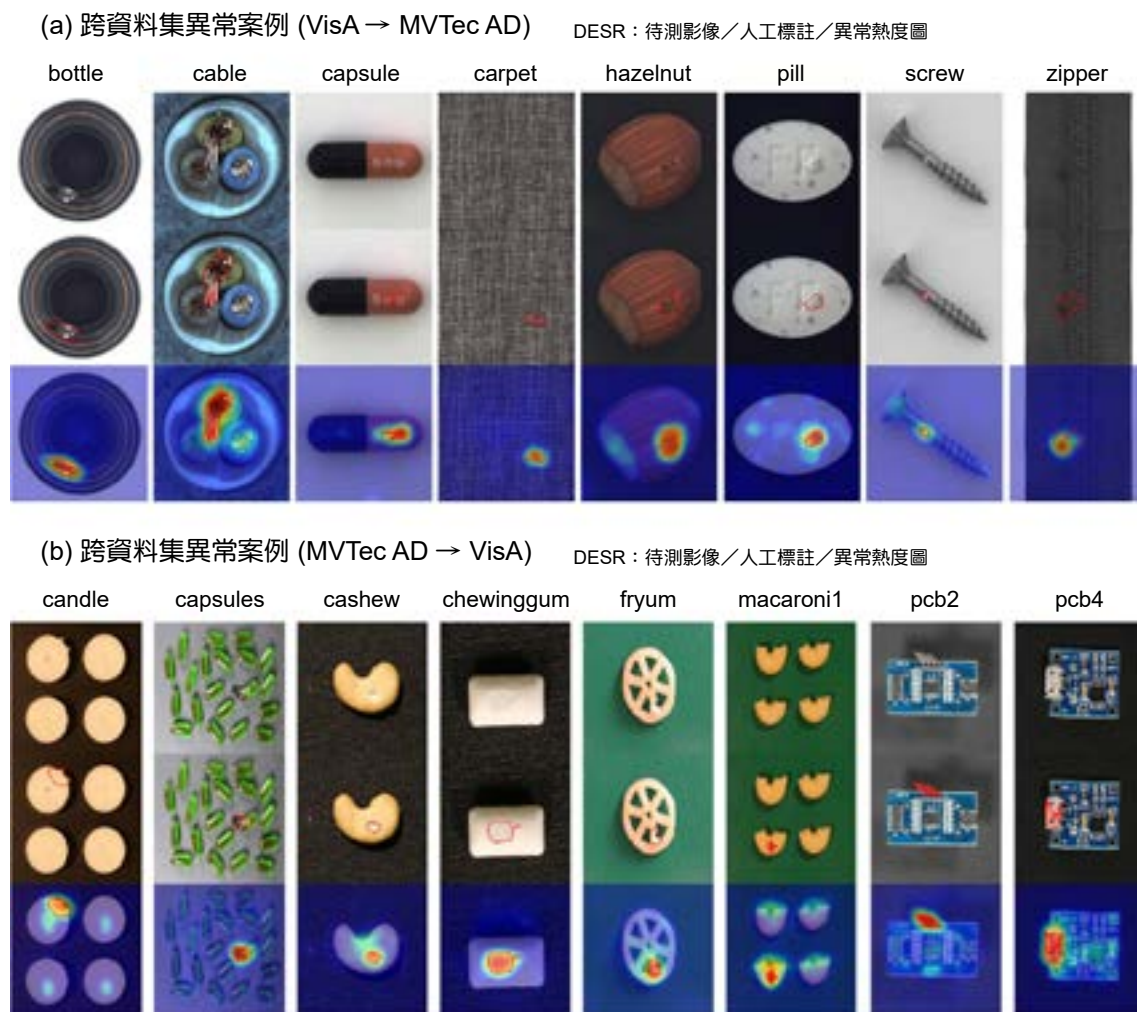


圖 6. 跨資料集異常偵測的真實案例。(a) 模型於 VisA 訓練、跨域測試於 MVTec AD 的結果；(b) 模型於 MVTec AD 訓練、跨域測試於 VisA 的結果。各欄由上而下為待測影像、人工標註的真實異常區域與模型異常熱度圖。影像取自作者團隊 DESR 研究⁽¹⁹⁾，重新排版。

4.4 幾何、多感測與生成先驗

單靠彩色影像，可能無法發現高度、內部結構或溫度異常。未來系統需整合深度、點雲、X 光、紅外線、聲音、振動與製程數值；由於不同模態在尺度、對位、缺失資料與取樣頻率上存在差異，單純串接特徵通常並不足夠。另一方面，語言控制的生成模型可依產品結構指定缺陷位置與型態，補足缺少的異常資料。真正有效的多模態與生成先驗，應在目標場域證明能降低誤報、提升小缺陷定位，並維持跨批次與跨設備的穩定性。

五、從異常偵測到異常理解

5.1 視覺語言模型讓檢測系統具備對話能力

傳統模型通常只輸出異常分數與熱度圖，缺陷類型及可能原因仍需由工程師自行判讀。視覺語言模型 (vision-language model, VLM) 則能同時理解影像與文字，因此使用者可以直接

用文字詢問「哪裡有異常？」或「這個缺陷可能是什麼？」。AnomalyGPT 將大型 VLM 導入工業檢測，使模型可回答異常是否存在、位於何處，並進行多輪對話⁽²⁰⁾。MMAD 則不是一個檢測模型，而是一套用來測試多模態大型模型是否真的理解工業異常的評估基準，涵蓋產品知識、異常辨識、定位與描述等七項能力⁽²¹⁾。作者團隊的知識－語言對齊研究，則進一步調整正常與異常的文字錨點，使語言概念與工業缺陷的視覺表徵對應得更穩定⁽¹⁹⁾。

5.2 異常理解必須連結可驗證證據

如同臨床診斷須指出病灶位置並說明診斷依據，理想系統應能回答五個問題：是否異常、異常是什麼、異常在哪裡、哪些影像證據支持判斷，以及應採取何種行動。以半導體為例，晶圓圖可設計成涵蓋缺陷存在、類型、位置、外觀與可能製程假設的多階段問答；在混合缺陷情境下，模型還必須輸出缺陷集合，並為每一類缺陷分別提供由影像計算而得的位置。模型提出的製程原因只能視為待驗證假設，而非已確認的物理根因。無論文字回答多流暢，或注意力是否落在正確位置，都不能單獨證明該位置真正決定了答案⁽²²⁾。

5.3 強化學習與結構化推理

近年方法開始採用群組相對策略最佳化 (group relative policy optimization, GRPO) 等強化學習技術。簡言之，GRPO 會使模型對同一問題產生多個候選回答，並以同一組候選的平均表現作為比較基準：表現優於群組平均的回答獲得強化，劣於平均者則被抑制。由於比較基準直接取自群組內部，不需再額外訓練一個價值評估模型，訓練流程相對簡單穩定。這種作法特別適合異常推理任務，因為此類任務的獎勵大多可由規則直接檢核——例如答案是否正確、框選位置是否與標註區域重疊、輸出格式是否完整——模型因此能在可驗證的回饋下，逐步保留較佳的推理與定位方式。AnomalyR1 代表從傳統監督式問答走向「同時推理與定位」的發展趨勢⁽²³⁾。不過，如果獎勵只看最後答案是否正確，模型仍可能靠類別比例、選項位置或固定模板猜答案，而沒有真正看懂影像。作者團隊近期提出的 EGVLR 因此把「看到了什麼證據」、「如何判斷」、「異常位於哪裡」與「最後答案」分開表示，再分別獎勵答案正確性、位置合理性與說明一致性；對正常影像則要求不要隨意框出缺陷。這種分階段訓練在 MMAD 上改善了需要視覺證據的推理與定位，也降低了無缺陷影像上的錯誤框選⁽²⁴⁾。圖 7 即為此類任務的實例：模型除須判定缺陷類型，亦須標示支持該判斷的影像位置，二者缺一皆不足以稱為真正的理解。

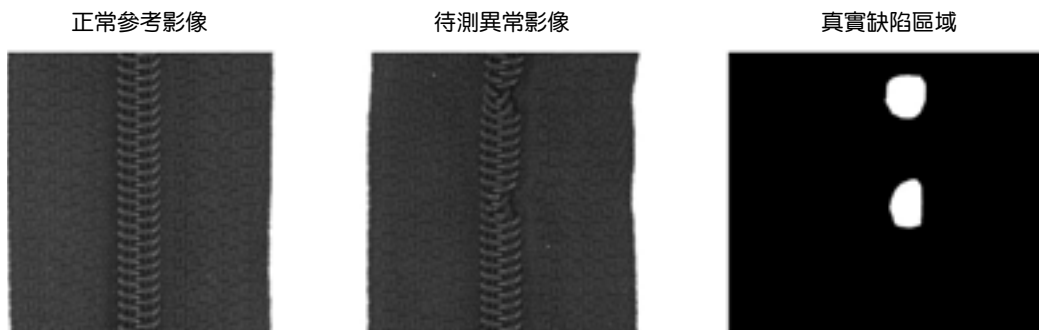


圖 7. 異常理解的實際案例。左為正常拉鍊，中為待測影像，可看到部分拉鍊齒被擠得更靠近；右為人工標註的真實缺陷區域。這類任務要求模型不只答對缺陷類型，也要指出支持判斷的位置。影像取自作者團隊 EGVLR 研究⁽²⁴⁾，重新排版。

5.4 代理式異常分析與主動取證

工程師面對難以判定的影像時，通常會放大可疑區域、尋找同型號的正常樣本、計算差異，或要求補充另一個視角。代理式人工智慧 (agentic AI) 希望讓模型也能依需要主動使用這些工具，而不是只看一次影像就立即下結論。ReAct是一種讓模型在「推理」與「採取行動」之間交替的通用框架⁽²⁵⁾；AnomalyClaw 則把這個概念用於異常檢測，讓模型透過多輪反駁、檢索與工具查證候選異常⁽²⁶⁾。然而，模型有呼叫工具，不代表它真的使用了工具提供的證據。多輪訓練仍需確認工具影像有進入模型的生成與評分流程，並透過遮蔽或打亂影像等測試，檢查答案是否真的會隨證據改變。

六、可信部署不能只看整體準確率

工業部署真正關心的是操作風險，而非單一排行榜分數。圖 8 整理四項應分開稽核的能力：任務準確度、視覺定位、因果視覺依賴與部署安全。同一模型可能在其中一項進步，另一項卻退步；高準確率、清楚的熱度圖、流暢的文字與完整的工具操作紀錄，都不能互相替代。可信評估還應納入跨階段一致性、端到端診斷正確性，以及不同訓練執行之間的穩定性。

可信賴異常理解：答案、證據、路徑與風險需分別稽核



圖 8. 可信賴的異常理解，需分別稽核任務準確度、視覺定位、因果依賴與部署安全。

6.1 誤報與漏檢具有不同成本

誤報 (false alarm, FA) 會讓正常產品進入複檢流程，降低產能或造成不必要的報廢；致命漏檢 (fatal miss, FM) 則可能使缺陷產品流出，影響客戶與安全。在高度不平衡的資料中，只要把所有產品都判為多數類別，仍可能獲得很高的整體準確率。例如在缺陷率為百分之一的產線上，一個一律判定為正常的模型即可達到百分之九十九的準確率，卻無法攔截任何缺陷品；若第一階段篩選錯誤，還可能觸發後續看似連貫、實則缺乏影像依據的類型、位置與製程說明。因此實際評估至少應報告 FA、FM、平衡準確率與固定誤報率下的召回率、信賴區間與 FA-FM 風險曲線。

這裡的「操作點 (operating point)」，是指模型實際上線時，因判定門檻設定而落在的那一組誤報率與漏檢率數值。同一個模型只要調整門檻，誤報率與漏檢率就會此消彼長，形成一條 FA-FM 曲線；門檻一旦選定，模型即固定在曲線上的某一點運作，這一點便是操作點。因此門檻不宜一律沿用 0.5，而應依產線的風險成本決定。

以一個簡化的例子說明。假設某產線導入初版模型後，量測得誤報率 12%、漏檢率 0.5%，複檢站因誤報過多而負荷過重；工程師加入抑制誤報的訓練介入後，誤報率降至 0.5%，漏檢率卻升至 3%。這並非單純的改善，而是操作點沿著曲線移動：誤報減少的同時，漏檢增加了六倍。若一件缺陷品流出的代價遠高於一次複檢，後者反而是更危險的設定。合理的作法，是先依誤報與漏檢各自的單位成本，在 FA-FM 圖上畫出可接受的範圍 (圖 9)，再檢查候選模型 (含以不同隨機種子重新訓練的結果) 是否都穩定落在此範圍內，而非只比較整體準確率。

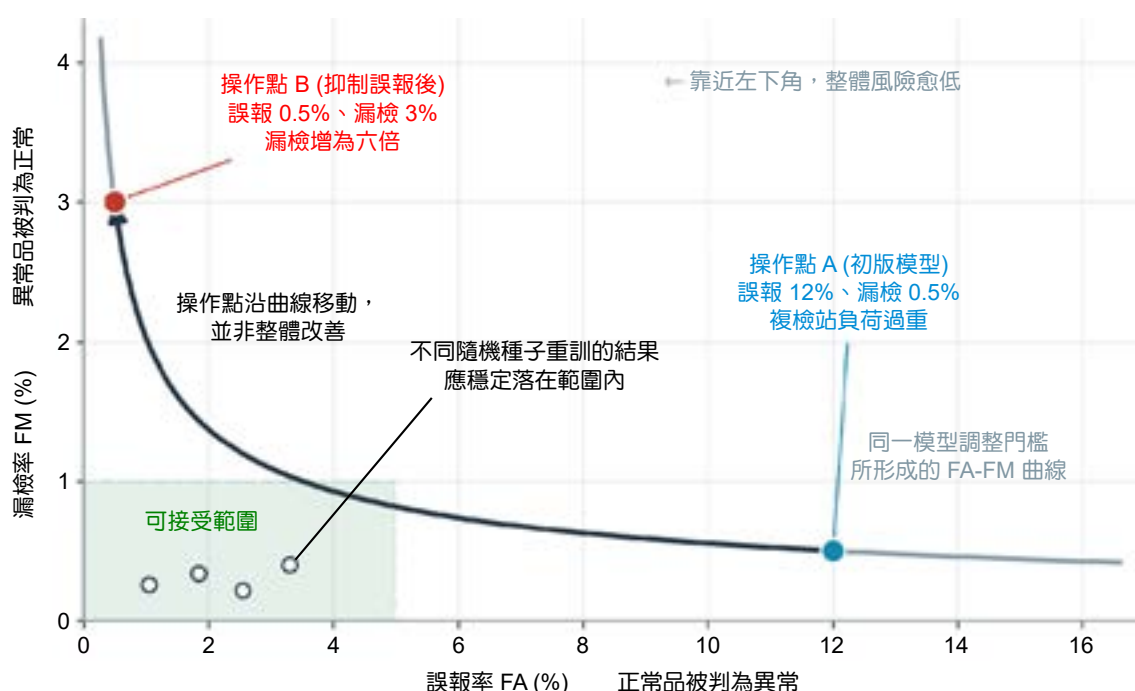


圖 9. FA-FM 風險平面與操作點示意 (數值為簡化範例，非實測結果)。同一模型調整判定門檻，誤報率與漏檢率即沿曲線此消彼長；由操作點 A 移動至 B 時，誤報大幅下降，漏檢卻增為六倍。可接受範圍依產線的誤報與漏檢單位成本劃定，候選模型 (含不同隨機種子的重訓結果) 應穩定落在範圍內。

6.2 影像層級與像素層級需分開評估

局部異常圖表現良好，不代表整張影像的判斷一定同樣可靠；反之，模型也可能正確判定產品為異常，但熱度圖卻集中在背景。常見的 AUROC (area under the ROC curve) 可理解為模型在不同判定門檻下區分正常與異常的整體能力；平均精確率則更能反映異常樣本較少時的排序品質，即依異常分數由高到低排列後，真正的異常樣本是否集中在前面。影像層級指標關心「整件產品判對了嗎」，像素層級與區域重疊指標則關心「缺陷位置找對了嗎」。因此，即使報告綜合分數，也應保留各項組成指標，避免平均值掩蓋定位或排序失敗。

6.3 因果視覺依賴與資料捷徑

本節要處理的問題，可類比為判斷學生是真正讀懂題目，還是只憑作答技巧猜中答案。單看分數無法分辨這兩種情形，必須改動題目本身，再看答案是否跟著改變。因果審計即依此邏輯進行：固定問題與正確答案，將正確答案直接提供給模型，讀取模型賦予該答案的機率作為基準，接著在其餘輸入完全不變的條件下更換影像，比較機率的變化。

常用的破壞方式有兩種。「遮蔽」是把影像換成空白，用以檢驗模型是否需要影像存在；「打亂」則是把像素在產品輪廓內重新排列，缺陷總量與外形維持不變、僅破壞空間排列，用以檢驗模型是否依賴缺陷的位置與形狀。若機率明顯下降，表示答案確實建立在影像證據之上；若幾乎不變，模型很可能不是看圖作答，而是依循文字端既有的統計規律——例如題目的問法通常對應哪一個答案，或固定的答題格式本身即足以推得答案，此即所謂的語言先驗 (language prior)。

要使審計具說服力，前後兩次測試應使用相同樣本與順序以進行配對比較，遮蔽與打亂的結果也須分開報告，並加入「不提供影像、僅給文字題目」等對照，界定流暢文字本身能證明的範圍。以晶圓圖問答為例，某模型在缺陷類型題上準確率極高，影像打亂後信心卻幾乎不變；加入反事實訓練資料——把打亂後的影像配上其實際對應的正確答案——重新訓練後，真實影像與打亂影像的信心差距才轉為顯著。惟審計至此仍未結束：視覺依賴改善不等於部署更安全，任何資料或損失函數的調整，都應回頭重新檢查各缺陷類別的 FA、FM 與多種子穩定性。

6.4 運算、資安與可追溯性

大型 VLM 的運算成本較高，高速產線可採分層式架構：先由小型模型進行篩選，再將困難案例交由大型模型比較與說明。例如產線每三秒產出一件產品時，小型模型可在數十毫秒內完成初判，僅將少數可疑案例送交大型模型，即可兼顧節拍與理解能力。製造影像涉及產品設計與製程機密，應優先考量廠內推論、權限控管、資料留存與模型版本管理。每次推論宜保存輸入影像、正常參考、提示與模型版本、門檻、熱度圖、工具紀錄、信心與人工覆核結果，形成可追溯的資料與決策鏈，以支援稽核、校準與持續改善。

七、智慧製造與跨領域應用

7.1 半導體與電子製造

半導體應用涵蓋晶圓圖、晶圓表面、積體電路顯微影像、印刷電路板與焊點檢測。異常可能是週期圖案中的局部破壞，也可能是多種缺陷同時存在。以晶圓圖為例，同一片晶圓可能同時出現中心圓斑、邊緣環狀與線性刮痕，各自對應不同製程站點；若模型只輸出單一類別，工程師仍須自行拆解混合訊號。未來系統可將缺陷存在、類型、位置、外觀與可能製程假設設計為多階段問答，並結合製程站點、設備參數、批次資訊與歷史缺陷地圖。模型產生的製程說明如 5.2 節所述，仍需搭配設備資料與工程師分析加以確認。

7.2 金屬、機械、組裝與包裝

金屬與機械零件需要檢測刮痕、裂縫、凹陷、加工不完整與尺寸變形；組裝檢測則關注元件的數量、方向與關係。食品、紡織與包裝產業可檢測異物、破損、印刷錯誤與封口不完整。二維影像適合辨識顏色與紋理差異，三維點雲與深度資訊可偵測幾何變形，多視角影像

則可降低遮蔽造成的漏檢。產品快速換線時，少樣本參考、統一模型與可調整的文字規格，有助於縮短導入時間。例如包裝產線每逢節慶改版，印刷圖樣與封膜材質常於短期內全面更換；若模型須重新蒐集數百張影像方能上線，改版初期便會出現檢測空窗。

7.3 醫療、遙測與設備監控

類似技術也可應用於核磁共振 (magnetic resonance imaging, MRI)、電腦斷層 (computed tomography, CT)、內視鏡、皮膚與病理影像，或延伸至衛星影像、熱像與設備聲音。跨領域測試可用來檢查模型是否真正學到通用的語意與結構；然而，不同領域在資料分布、法規與錯誤成本上差異極大。工業資料上的良好結果，不能直接等同於具備臨床或公共安全部署資格；仍需經過領域專家驗證、獨立資料測試、校準與風險管理。

八、未來發展方向

第一，模型將由「每項產品各自訓練」，走向可依正常參考、文字規格與感測條件動態適應的通用異常系統。第二，語意模型與稠密結構模型將更緊密整合，讓「理解異常概念」與「精準定位微小異常」同時成立。第三，影像、點雲、熱像、聲音與製程數值將形成多感測閉迴路，模型可在資訊不足時主動要求額外觀測。第四，異常理解將與智慧代理及具身巡檢結合，依序完成感知、比較、反駁、查詢、決策與行動。第五，可信賴評估將從平均準確率，擴展至因果證據、FA-FM 風險曲線、校準、長期漂移、跨批次與多種子穩定性。

真正的技術進步，不只是在既有資料集上增加少數分數，而在於降低新產品導入成本、減少誤報與漏檢、即使環境改變仍能維持可靠，並讓工程師能檢查模型究竟使用了哪些證據。異常偵測也將從獨立演算法，轉型為串接感測器、資料平台、製程知識與人員決策的品質工程系統。

九、結論

異常偵測技術已從人工規則，逐步發展至正常資料重建、知識蒸餾、特徵記憶庫、合成異常、少樣本與零樣本方法。視覺基礎模型提供可遷移的語意與結構知識，使統一模型、跨域檢測與多模態分析日益可行；視覺語言模型則讓系統能描述異常、回答問題、比較正常參考，並提出診斷假設。

然而，新一代系統的價值，不在於能否說得像專家，而在於其判斷是否建立在正確、可驗證且可追溯的視覺證據上。未來的智慧檢測必須同時處理任務準確度、空間定位、因果視覺依賴、工具證據是否真正被使用、誤報與漏檢、運算效率、資安及長期穩定性。換言之，異常分析的下一階段，不只是讓機器看見缺陷，而是讓機器理解差異、主動查證，並以可稽核且安全的方式協助人類採取行動。

參考文獻

1. Y. Cao, X. Xu, J. Zhang, Y. Cheng, X. Huang, G. Pang, and W. Shen, arXiv:2401.16402 (2024).
2. X. Jiang, B. Hu, and F. Zheng, “A Survey of Recent Advances in Industrial Anomaly Detection: From Normal-Only Training to Foundation-Model Priors”, Preprints, 202606.0719 (2026).
3. Y.-H. Hsieh and S.-H. Lai, “CSAD: Unsupervised Component Segmentation for Logical Anomaly Detection”, *British*

- Machine Vision Conference*, November 25-28, (2024).
4. P. Bergmann, M. Fauser, D. Sattlegger, and C. Steger, "MVTec AD-A Comprehensive Real-World Dataset for Unsupervised Anomaly Detection", *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 16-20, (2019).
 5. K. Roth, L. Pemula, J. Zepeda, B. Scholkopf, T. Brox, and P. Gehler, "Towards Total Recall in Industrial Anomaly Detection", *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 18-24 (2022).
 6. V. Zavrtanik, M. Kristan, and D. Skočaj, "DRAEM-A Discriminatively Trained Reconstruction Embedding for Surface Anomaly Detection", *IEEE/CVF International Conference on Computer Vision*, October 11-17, (2021).
 7. S.-C. Lin and S.-H. Lai, "SQUAD: Scalar Quantized Representation Learning for Unsupervised Anomaly Detection and Localization", *European Conference on Computer Vision Workshops*, September 29 - October 4, 166 (2024).
 8. S.-C. Lin, H.-W. Lee, Y.-H. Hsieh, C.-Y. Ho, and S.-H. Lai, "Masked Attention ConvNeXt UNet with Multi-Synthesis Dynamic Weighting for Anomaly Detection and Localization", *British Machine Vision Conference*, November 20-24 (2023).
 9. S.-C. Lin and S.-H. Lai, "LFQUIAD: Lookup-Free Quantized Autoencoder for Few-Shot Unsupervised Industrial Anomaly Detection via Synthetic Diffusion Inpainting", *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, June 11-12, (2025).
 10. A. Radford, J. W. Kim, C. Hallacy, et al., "Learning Transferable Visual Models from Natural Language Supervision", *International Conference on Machine Learning*, July 18-24, (2021).
 11. M. Caron, H. Touvron, I. Misra, et al., "Emerging Properties in Self-Supervised Vision Transformers", *IEEE/CVF International Conference on Computer Vision*, October 11-17, (2021).
 12. J. Jeong, Y. Zou, T. Kim, D. Zhang, A. Ravichandran, and O. Dabeer, "WinCLIP: Zero-/Few-Shot Anomaly Classification and Segmentation", *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 18-22, (2023).
 13. Q. Zhou, G. Pang, Y. Tian, S. He, and J. Chen, "AnomalyCLIP: Object-Agnostic Prompt Learning for Zero-Shot Anomaly Detection", *International Conference on Learning Representations*, May 7-11, (2024).
 14. S.-C. Lin and S.-H. Lai, "CLIP-FSQAE: CLIP-Guided Finite Scalar Quantized Autoencoder for Few-Shot Anomaly Detection", *IEEE International Conference on Image Processing*, September 14-17, 313 (2025).
 15. S.-C. Lin, J.-N. Hwang, and S.-H. Lai, "BSMAD: Bridging Semantic and Structural Manifolds for Robust Cross-Modality Medical Anomaly Detection", *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, June 3-4, (2026).
 16. Z. You, L. Cui, Y. Shen, K. Yang, X. Lu, Y. Zheng, and X. Le, "A Unified Model for Multi-Class Anomaly Detection", *Advances in Neural Information Processing Systems*, November 28 – December 9, (2022).
 17. S.-C. Lin and S.-H. Lai, "UniLEAD: A Unified and Lightweight Model for Anomaly Detection", *IEEE/CVF International Conference on Computer Vision Workshops*, October 19-20, 1483 (2025).
 18. S.-C. Lin, J.-N. Hwang, and S.-H. Lai, "UFMADS: A Unified Framework for Prompt-Guided Medical Anomaly Detection and Segmentation", *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, June 3-4, 1291 (2026).
 19. S.-C. Lin, "DESR: Dual-Encoder Semantic Retrieval for Open-Set Anomaly Discovery", *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD)* (2026).
 20. Z. Gu, B. Zhu, G. Zhu, Y. Chen, M. Tang, and J. Wang, *Proceedings of the AAAI Conference on Artificial Intelligence*, **38** (3), 1932 (2024).
 21. X. Jiang, J. Li, H. Deng, et al., "MMAD: A Comprehensive Benchmark for Multimodal Large Language Models in Industrial Anomaly Detection", *International Conference on Learning Representations*, April 24-28 (2025).
 22. S. Jain and B. C. Wallace, "Attention is not Explanation", *Conference of the North American Chapter of the Association for Computational Linguistics*, June 2-7 (2019).
 23. Y. Chao, J. Liu, J. Tang, and G. Wu, "AnomalyR1: A GRPO-Based End-to-End MLLM for Industrial Anomaly Detection", arXiv:2504.11914 (2025).
 24. S.-C. Lin, Y.-H. Lu, D. Y. Ye, and S.-H. Lai, "EGVLR: Evidence-Grounded Vision-Language Reinforcement for Anomaly Reasoning", *European Conference on Computer Vision*, (2026).
 25. S. Yao, J. Zhao, D. Yu, et al., "ReAct: Synergizing Reasoning and Acting in Language Models", *International Conference on Learning Representations*, May 1-5, (2023).
 26. X. Jiang, Y. Zhao, Z. Yang, and F. Zheng, "AnomalyClaw: A Universal Visual Anomaly Detection Agent via Tool-Grounded Refutation", arXiv:2605.10397, (2026).

作者簡介

林士智先生為國立清華大學跨院國際博士班學位學程 博士候選人。

Shih-Chih Lin is currently a Ph.D. candidate in the International Intercollegiate Ph.D. Program at National Tsing Hua University.

賴尚宏先生為美國佛羅里達大學電機與電腦工程博士，現職為國立清華大學資訊工程學系教授兼電機資訊學院副院長。

Shang-Hong Lai received his Ph.D. in Electrical and Computer Engineering from the University of Florida. He is currently a professor in the Department of Computer Science and Associate Dean of the College of Electrical Engineering and Computer Science at National Tsing Hua University.

利用微生理類器官微流體系統建立肝臟安全性及代謝模型

Building a Liver Safety and Metabolism Model by MPS Organoid Microfluidic Systems

趙柔雅、余昊翰、洪明秀、曾繁根

Rou-Ya Zhao, Hao-Han Yu, Ming-Shiu Hung, Fan-Gang Tseng

隨著科技的快速發展與倫理意識的提升，全球對動物實驗替代方法的需求日益增加。本研究發展一套可重組式模組化多器官晶片平台，以單一雙腔室晶片單元為基礎，透過上下獨立微流道與多孔膜分隔培養區，支援靜態培養及雙向動態灌流；多個晶片單元可依需求組成線性或矩陣式陣列，並透過流體串接建立多器官系統。研究以肝臟與腫瘤晶片為示範，建立肝臟-腫瘤器官晶片模型，探討肝臟代謝、腫瘤負荷與抗癌藥物反應之間的交互作用。結果顯示，雙腔室多孔膜架構可提升肝細胞尿素生成，而調整肝臟與腫瘤細胞比例則可改變 Regorafenib (REG) 的藥效與肝細胞毒性。此平台具備可重組、可擴充及可調整器官比例等特色，可作為多器官藥物評估與精準用藥研究之微生理系統。

With the rapid advancement of technology and the growing need for alternatives to animal testing, this study developed a reconfigurable modular multi-organ-on-chip platform based on a bicompartmental chip unit. The unit contains independent upper and lower microfluidic channels separated by a porous membrane and supports both static culture and bidirectional dynamic perfusion. Multiple units can be assembled into linear or matrix arrays and interconnected to construct customised multi-organ systems. As a proof of concept, liver and tumour chips were integrated to investigate the interactions among hepatic metabolism, tumour burden, and anticancer drug responses. The results showed that the dual-chamber membrane architecture enhanced hepatic urea secretion, while varying the liver-to-tumour cell ratio altered the efficacy of Regorafenib (REG) and hepatotoxic responses. The modular platform therefore provides a reconfigurable and scalable microphysiological system for multi-organ drug evaluation and precision medicine research.

一、組合式器官晶片應用於細胞培養

1.1 器官晶片應用於細胞培養之關聯性

微生理系統 (microphysiological systems, MPS) 是一種整合微流體技術、組織工程與細胞

生物學的創新平台，可在體外模擬人類器官的生理功能，為動物實驗提供高效且具倫理性的替代方案。在臺灣，微流體技術與生醫晶片的發展逐漸成熟，尤其是體外肝臟模型作為藥物篩選與毒性評估的重要平台，契合了 3R 政策的核心理念，成為藥物開發與毒性測試領域的研究重點。

相較於動物模型，器官晶片平台具有明顯的優勢。動物模型雖可進行整體性研究，支持長期深入的臨床轉化研究，並擁有豐富的既有數據，但也存在物種差異性、倫理道德顧慮、昂貴成本與有限的人類相關性等不足。器官晶片通過模擬人體微環境，能顯著提高模型的人類相關性，甚至構建個體化模型。同時，它可用於構建高通量篩選平台，降低動物模型的研發費用，並在生醫藥研究中實現減量和替代動物實驗的目標。此外，器官晶片在藥物篩選中已成功用於毒性事件的預測，例如心血管、肝臟或皮膚毒性，並能識別脫靶效應，為藥物重新設計提供依據，從而減少對患者的潛在風險並降低開發成本。尤其是具有微流體的多器官脈管系統，能適用於標靶藥物開發中脫靶效應的臨床前毒性測定，顯示出巨大的應用潛力。

1.2 利用組合式器官晶片作為培養細胞之平台：

本研究以 Transwell 之雙腔室概念為基礎，結合微流體技術建構上下獨立流道與多孔膜分隔之單元，以克服傳統 Transwell 僅能進行靜態培養、無法重現體內動態流體剪切力之限制，並可依實驗需求以「mix-and-match」方式組合成線性陣列或矩陣陣列，進一步建立不同器官組合及器官比例的多器官晶片系統。

透過上下獨立流道與持續灌流系統，使培養液能穩定流經細胞培養區域，模擬體內血液循環與組織間的物質傳輸過程，進而提供更具生理相關性的細胞生長環境，提升體外模型對器官功能與藥物反應的模擬能力。為能夠自由組合型的晶片，在操作時可依不同需求搭配不同數量的晶片並予以組裝結合。單一晶片由上晶片與下晶片構成，上晶片下方以熱壓方式貼合 PETE 多孔膜，此多孔膜之目的在於培養器官組織，提供交換物質之環境，並可同時透過上、下晶片的流道分別接觸到培養液。於使用時在四個角落，從下晶片往上晶片方向以平口螺絲栓緊，由於上晶片下側之外圍鑲嵌有 O-ring，螺絲栓緊的同時也可侷限下晶片流道空間的液體，如圖 1 淺藍色所示。於此同時，下晶片流道可透過側邊的銜接孔位與其他晶片進行橫向串聯，讓橫向流道內之液體同時流通 PETE 的下方(下晶片流道空間)。同理，如圖 1 淺橘色所示，上晶片流道可透過側邊的銜接孔位與其他晶片進行縱向串聯，讓縱向流道內之液體同時流通 PETE 的上方(上晶片流道空間)。晶片組裝完成之後，可於四周的銜接孔位扣上彈性皮套避免液體洩漏，並於正上方壓緊 clamp 用以注入流體(如圖 1 淺灰色所示)。晶片銜接孔位埠會套上合適的 O-ring (藍色小點) 以避免組裝介面的洩漏。

1.3 實驗

使用組合式器官晶片作為雙層細胞 (HepG2/C3A 及 HUVEC-sVAP-1) 的培養支架，先分別在多孔 PET 膜兩側接種 HUVEC-sVAP-1 與 HepG2/C3A，形成雙層細胞結構。細胞接種前分別以 CellTracker™ Green CMTPX Dye 及 CellTracker™ Red CMTPX Dye 標記，並使用 Confocal microscope LSM-780 觀察。結果顯示兩種細胞均可穩定附著於多孔膜上下兩側，證實此模組化晶片可作為雙側細胞培養與器官晶片建構平台。

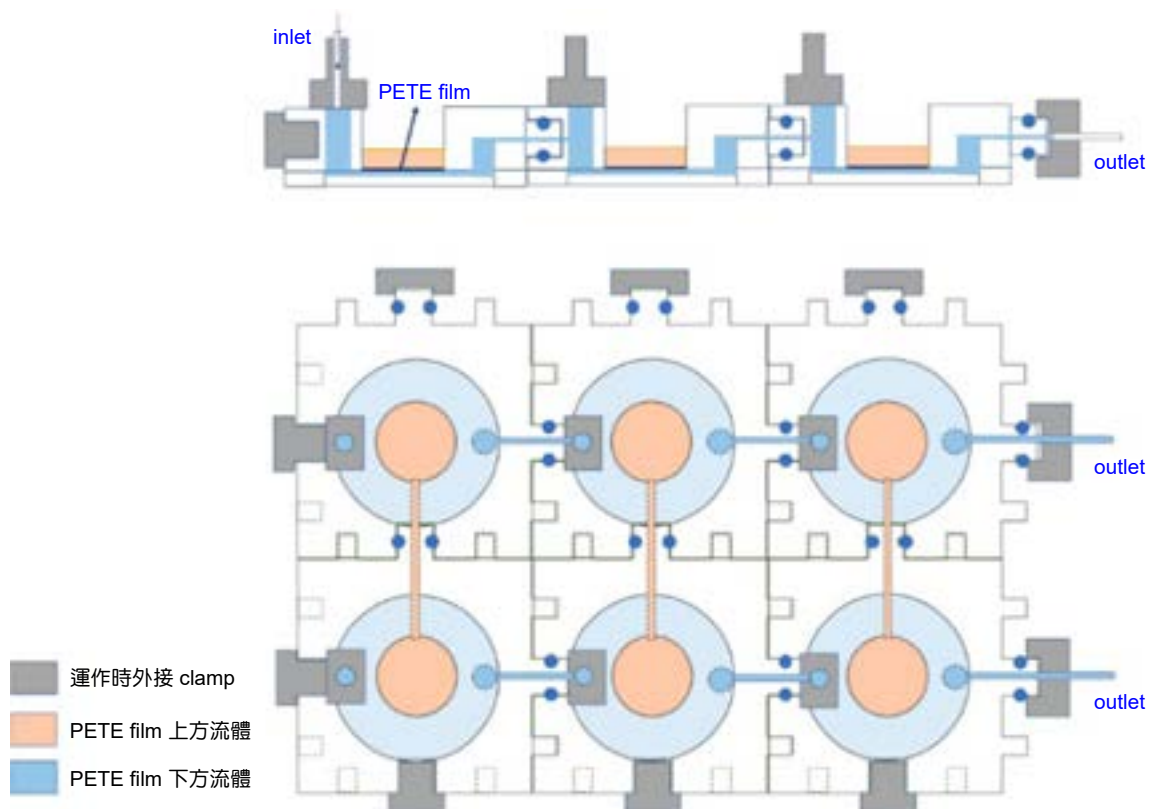


圖 1. 組合式器官晶片概念示意圖。

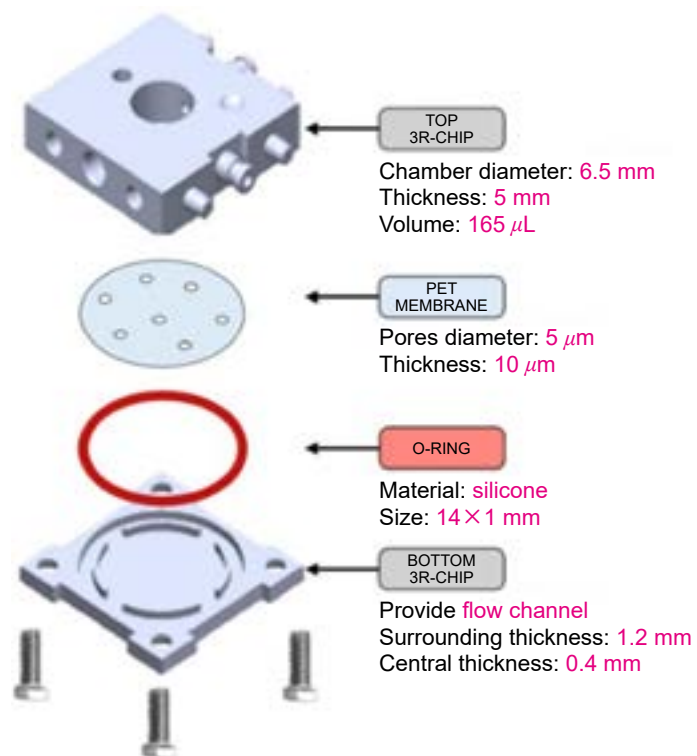


圖 2. 組合式器官晶片裝置系統組成示意圖。

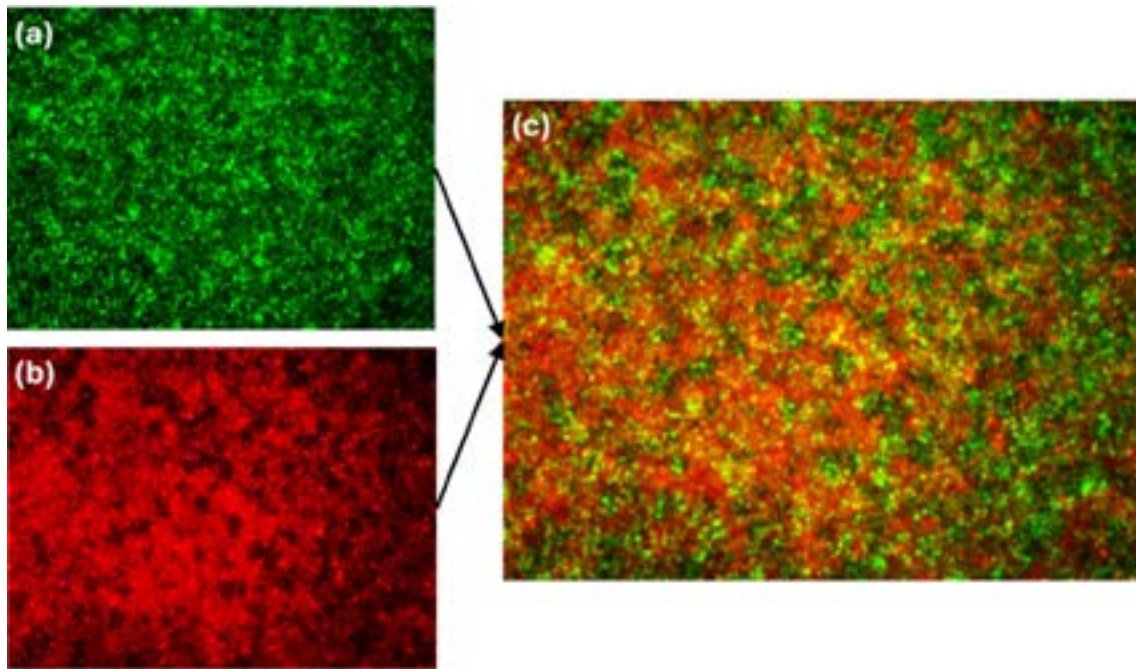


圖 3. 使用螢光顯微鏡拍攝組合式器官晶片之雙層細胞結構。(a) HUVEC-sVAP-1 (b) HepG2/C3A 及 (c) 雙層細胞結構。

二、組合式器官晶片應用於肝臟代謝功能檢測

2.1 微流控細胞培養平台與肝臟代謝功能檢測之關聯性

肝臟為人體維持代謝解毒、蛋白質合成與體內平衡之核心器官。於臨床醫學與藥物評估中，通常以血清白蛋白的分泌水平以及尿素循環的運作效率，作為評估肝細胞合成解毒功能的重要功能性指標。在慢性肝病或受損的肝細胞中，白蛋白合成量常因功能受損而顯著降低⁽¹⁾，且其分子結構之改變亦會導致運輸能力衰退；而尿素循環亦會因功能受損而受阻，無法有效將具備細胞毒性的體內氨轉化為尿素，進而引發血氨過高及肝性腦病等嚴重併發症⁽²⁾。因此，精準監測此兩項代謝指標之變動，對於肝臟疾病診斷與藥物毒性管理具有關鍵意義。

然而，傳統體外平面培養模型往往缺乏體內微環境的動態刺激。先前的研究中證實，透過微流控細胞培養平台能高度模擬肝臟的複雜細胞結構與生理微環境⁽³⁾。此類器官晶片不僅能精準再現內皮細胞的窗孔結構 (fenestrae) 以及血液與肝細胞間的動態物質交換，亦能顯著提升肝細胞分泌白蛋白、合成尿素及細胞代謝酶 (如 CYP450) 之活性表現⁽³⁾。此外，在細胞免疫反應研究中，該平台亦能高精度重現中性白血球 (neutrophils) 從捕獲、附著到在內皮層爬行之完整招募響應行為，成為探討肝發炎或纖維化等病理機轉的理想體外平台⁽³⁾。

因此，利用微流控細胞培養平台建構仿生動態環境，能有效進行肝臟代謝功能之高效檢測，進而精準評估外源性藥物毒性所誘發之肝損傷程度。

2.2 利用 HepG2/C3A 細胞與 HUVEC-sVAP-1 作為仿生肝臟

在組合式器官晶片中將人類肝癌細胞 HepG2/C3A 6 萬顆接種於上層，並將人類臍靜脈內皮細胞 HUVEC-sVAP-1 5 萬顆接種於下層，靜置培養至貼附穩定並形成單層結構後，分別於靜態條件下進行未加藥、5-Fluorouracil (5-FU) 與 Regorafenib (REG) 處理，並培養 72 小

時，觀察尿素 (Urea) 合成與 CYP3A4 活性。另以傳統 96 孔盤作為比較平台。研究進一步顯示，在無藥物條件下，模組化晶片與 96 孔盤的細胞增殖、存活率及 CYP3A4 活性大致相當，但晶片中的尿素分泌量顯著較高，顯示雙腔室與多孔膜所形成的雙側物質交換環境可能有助於維持肝細胞合成功能。⁽⁴⁾

2.3 實驗

2.3.1 尿素檢測

在培養 72 小時時使用 Urea Assay Kit II (Abcam) 測量不同細胞數及藥物組別的細胞上清液 Urea 濃度。如圖 4 所示，未加藥組具有最高的 Urea 產量，約為 50 μg ；5-FU 組約下降至 25 μg ；REG 組約為 40 μg 。相較於傳統 96 孔盤，組合式晶片之 Urea 表現較高。根據研究，作者推測此差異可能與雙腔室架構提供雙側物質傳輸、較大的下層培養液儲存空間及較短的氧氣擴散距離有關；然而，細胞極化與尿素生成改善之間的直接關係仍需透過極性蛋白標誌與轉錄體分析進一步驗證。⁽⁴⁾

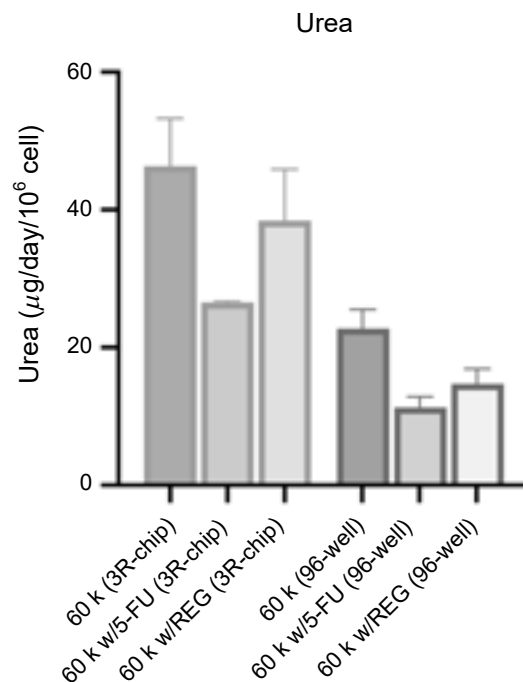


圖 4. HepG2/C3A 細胞於不同培養平台及藥物處理下之尿素分泌表現量。

2.3.2 細胞代謝酶 CYP3A4 之活性表現

為了解不同培養平台對 HepG2/C3A 細胞 CYP3A4 活性的影響，使用 P450-Glo™ CYP3A4 Assay Kit (Promega) 進行分析。實驗以 6 萬顆細胞培養 72 小時後，測定細胞 CYP3A4 活性，並比較模組化晶片與傳統 96 孔盤之間的差異。

由實驗結果可知，模組化晶片與傳統培養孔板的細胞增殖情形相近圖 5，且兩種平台的細胞存活率與 CYP3A4 活性亦無顯著差異。然而，肝細胞在模組化晶片中的尿素分泌量則顯著高於傳統培養孔板，顯示兩種平台雖具有相近的細胞生長、存活及基礎 CYP3A4 代謝能力，但模組化晶片可能更有利於部分代謝功能的維持。

由於本實驗是在靜態培養條件下進行，且兩種平台上層培養區的培養液體積相同，因此尿素分泌量的差異可能與模組化晶片的雙腔室結構及多孔膜介面有關。在傳統 96 孔盤中，細胞直接培養於不具通透性的塑膠表面，營養物質供應及代謝物交換主要侷限於細胞的頂端側；相較之下，模組化晶片利用多孔膜作為上下腔室之間的介面，使細胞可進行雙側的物質傳輸與擴散。

此外，下方腔室最多可提供 250 μL 的培養液空間，不僅增加細胞可利用的整體營養來源，也可作為代謝副產物的額外容納空間。此種雙腔室配置可能有助於降低局部代謝物累積所造成的抑制作用，進而維持較高的尿素循環活性。另一方面，多孔膜亦可縮短下方腔室氧氣向細胞擴散的距離，可能改善傳統靜態 96 孔盤深孔結構所造成的氧氣供應限制。

整體而言，模組化晶片與傳統培養孔板具有相近的 CYP3A4 活性，顯示兩種平台的基礎異生物質代謝能力並無明顯差異；然而，模組化晶片中的尿素分泌量顯著提升，則顯示其雙腔室與多孔膜結構可能為肝細胞的代謝與生合成作用提供較佳的微環境。

未來可進一步導入細胞極性相關的頂端側／基底外側 (apical/basolateral) 蛋白質標誌，以及轉錄體分析，以更明確地確認細胞於此微環境中的極化程度，並進一步釐清模組化晶片提升尿素生成能力的可能原因。⁽⁴⁾

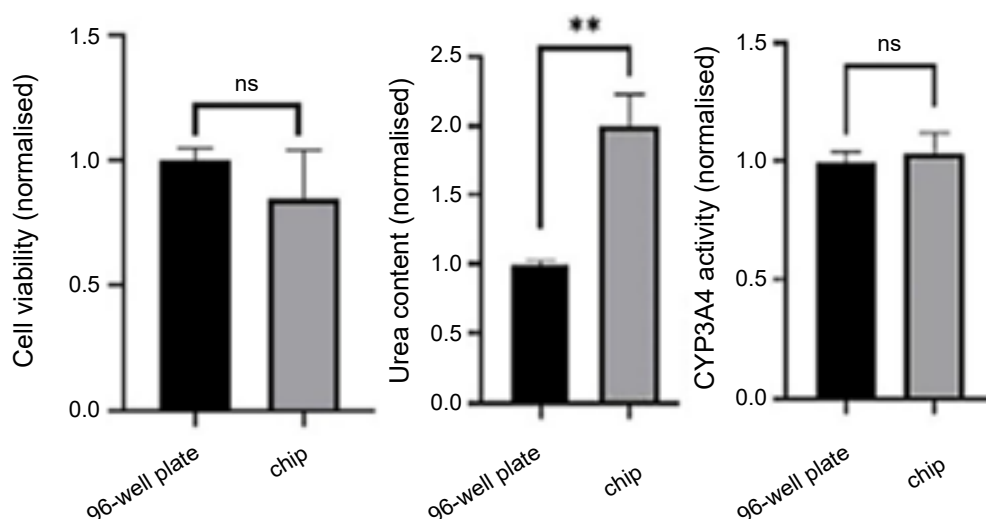


圖 5. HepG2/C3A 細胞於 96 孔盤與器官晶片平台之細胞存活率、尿素生成量及 CYP3A4 活性比較。

三、組合式器官晶片應用於腫瘤藥物反應之調控

3.1 微流控細胞培養平台與腫瘤藥物反應調控之關聯性

生物物理環境是腫瘤微環境 (tumor microenvironment, TME) 中的重要組成部分之一，包含了剪應力 (shear stress)、壓力梯度與流體流動等物理刺激。隨著微流體與器官晶片技術的發展，研究人員得以於體外模擬真實的血流與剪力環境，進一步觀察其對腫瘤細胞遷徙、轉移與藥物反應的影響，與傳統靜態細胞培養系統相比，動態微流體晶片具備可控流體灌流與剪力模擬能力，可更真實反映體內血管環境⁽⁴⁾。

3.2 利用 HT29 與 HUVEC 細胞作為腫瘤晶片與肝臟晶片串接

在組合式器官晶片中將人類結直腸癌細胞 HT-29 接種於上層，並將 HUVEC 內皮細胞接種於下層，靜置培養至貼附穩定後，組裝晶片並接入蠕動幫浦驅動的流體循環系統，於動態灌流條件下進行觀察。模組化晶片可在垂直與水平方向串接，使上、下流道分別傳遞流體。以螢光微粒追蹤流場，於 0.3 rpm、2 mm 內徑管路條件下，量測晶片不同位置之流速約為 114.5-241.65 $\mu\text{m/s}$ ；入口與出口流速較高，中央區域較低。依實測流速估算，下層腔室細胞可受到接近靜脈生理範圍的剪應力，顯示此平台具有提供生理相關動態流體環境的潛力。⁽⁴⁾

3.3 實驗

在組合式器官晶片中將 5 萬顆 HT-29 細胞接種於上層，並將 HUVEC 內皮細胞接種於下層，靜置培養至貼附穩定後，組裝晶片並接入蠕動幫浦驅動的流體循環系統，加入 Regorafenib (REG) 進行動態灌流，為了在多器官串接實驗中建立精準之細胞比例，調整肝臟細胞接種量為 5 萬顆。本研究固定每個肝臟晶片為 5 萬顆 HepG2/C3A，並將 HT-29 細胞數調整為 6 萬、5 萬、2.5 萬及 1.6 萬顆，建立四種肝臟：腫瘤細胞比例，分別為 1:1.2、1:1、1:0.5 及 1:0.32。以 30 μM REG 動態灌流 24 小時後，利用 CCK-8 評估肝細胞與腫瘤細胞存活率，並以 P450-Glo™ CYP3A4 assay 評估肝細胞代謝功能。結果顯示，腫瘤負荷與 REG 處理會共同影響肝臟 CYP3A4 活性、肝細胞存活率及腫瘤細胞存活率。

四、結論

本研究以組合式器官晶片平台建構肝臟代謝功能模型以及腫瘤模型，並將兩者串接，模擬藥物經肝臟代謝後對腫瘤細胞作用的變化，並評估肝臟雙向交互作用對藥物反應的調節影響。此平台的核心特色在於單一雙腔室晶片單元可支援靜態培養與雙向動態灌流，並能透過模組化串接建立線性或矩陣式多器官系統，進一步調整不同器官或組織之比例。肝臟晶片以 HepG2/C3A 與 HUVEC-sVAP-1 建構，結果顯示其具備尿素與 CYP3A4 等基礎肝臟功能。

在多器官串接實驗中，肝臟與腫瘤細胞比例被設計為 1:1.2、1:1、1:0.5 及 1:0.32 四種條件，並以 30 μM REG 動態灌流 24 小時。結果顯示，未加藥時，肝臟：腫瘤比例越高 (亦即腫瘤細胞越少) CYP3A4 活性越高；加入 REG 後，高腫瘤負荷組 (1:1.2 與 1:1) 的 CYP3A4 活性相較未加藥組顯著上升，而低腫瘤負荷組 (1:0.5 與 1:0.32) 差異較小。同時，REG 處理下腫瘤細胞數增加會伴隨肝細胞存活率下降，顯示器官比例會影響藥物反應。作者推測，高腫瘤負荷下的營養競爭、腫瘤分泌因子或局部藥物分布改變可能參與 CYP3A4 誘導，但目前尚未透過細胞激素分析或 LC-MS/MS 代謝體分析直接證實。

未來，本研究的仿生肝臟晶片將進一步測試更多肝功能性指標，例如 Albumin、UGT1A1/UGT1A9 等，以更完整評估肝臟功能。針對雙腔室架構所造成之尿素生成提升，亦可透過 ZO-1、MRP2 等極性相關蛋白染色及轉錄體分析，進一步確認細胞極化與功能改善之關係。

在肝臟－腫瘤串接模型中，將使用 LC-MS/MS 分析實驗後培養基中的 REG、M-2、M-5 等成分，直接驗證肝臟代謝與腫瘤藥物反應之關係。Analyst 指出 REG 可經 CYP3A4

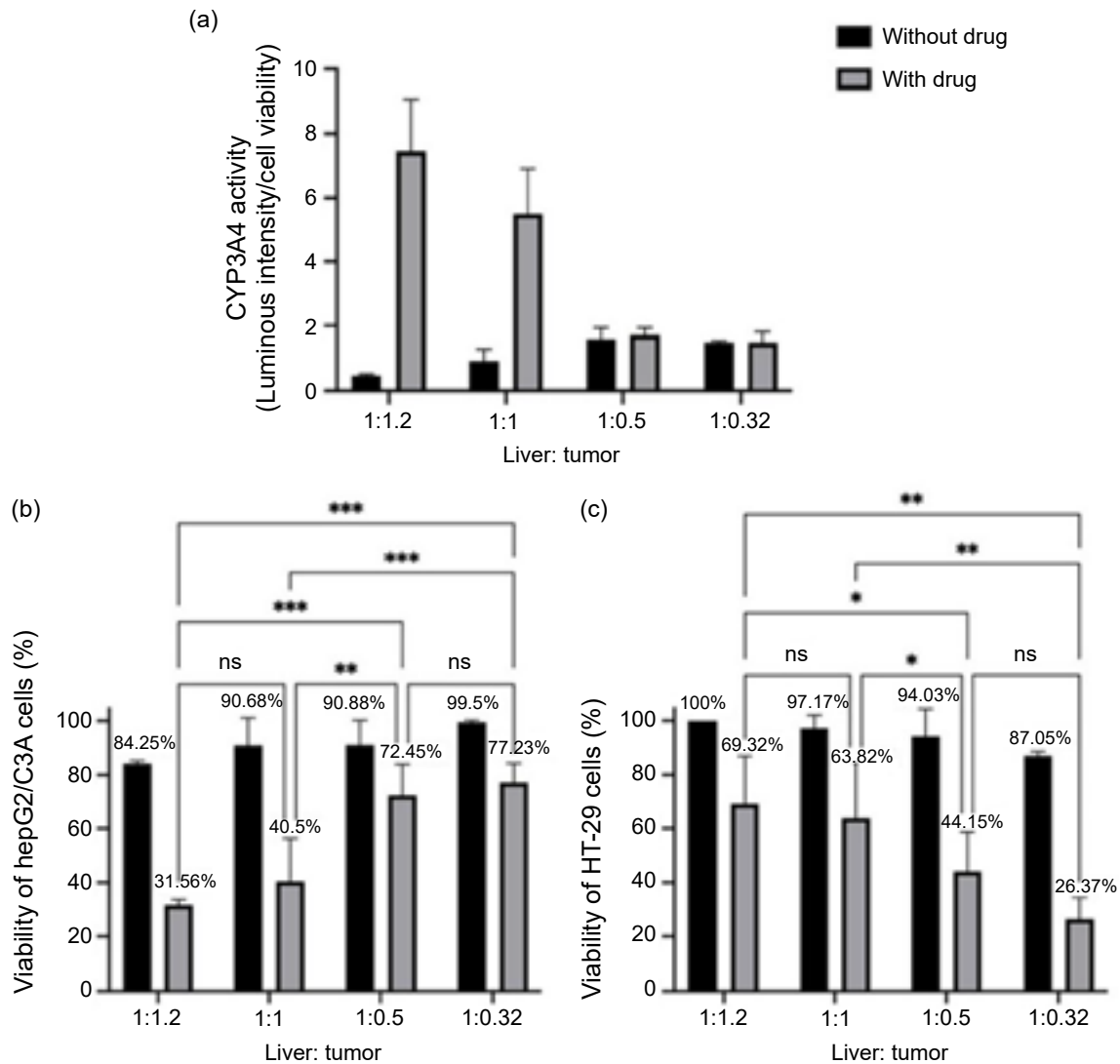


圖 6. 肝臟－腫瘤串接晶片於不同細胞比例下之藥物反應與存活率分析 (a) 肝細胞 CYP3A4 活性分析 (b) 肝細胞存活率 (c) 腫瘤細胞存活率。

代謝形成具有藥理活性的代謝物，但目前相關推論仍主要來自細胞存活率與 CYP3A4 等終點分析，尚缺乏直接代謝體證據。此外，可導入薄膜電化學感測器或奈米感測器，進行 Albumin、CYP 活性與藥物濃度的即時、非破壞性監測，並進一步將晶片出口與 LC-MS/MS 串接，建立線上代謝物分析系統。

基於組合式器官晶片的多器官串接特點，未來將進一步增加腎臟晶片，建立肝臟－腎臟及肝臟－腎臟－腫瘤等複合模型，並可延伸至 gut-liver、tumour-immune 等更複雜的器官交互作用。同時，可透過微結構培養腔促進 3D spheroid 或 organoid 形成，使模型更接近真實組織；並系統性分析與最佳化灌流參數及剪應力，以提升平台在藥物動力學、毒性預測與精準藥物篩選上的應用價值。

五、誌謝

本研究感謝國家科學及技術委員會 (NSTC) 3R 計畫 (NSTC 114-2321-B-007-001)、國家衛生研究院 (NHRI) 晶創計畫 (115-1401-01-30-04、114-1401-01-30-05)，以及國立清華大學－國家衛生研究院合作計畫「國立清華大學種子計畫」(112F7MDIE1)，並獲教育部高等教育深耕計畫支持。

參考文獻

1. Nijin Wu *et al.*, *Molecular Medicine Reports*, **29** (2), (2024).
2. Wang, H., J. Ran, and T. Jiang, *Subcell Biochem*, **73** (7), 7 (2014).
3. Xiaojun Zhang *et al.*, *Cancers* (Basel), **14** (3), 348 (2022).
4. Yun-Jie Hao *et al.*, *Analyst*, **151** (16), 4628 (2026).

作者簡介

趙柔雅女士現為國立清華大學工程與系統科學系碩士生。

Rou-Ya Zhao is currently a M.S. student in the Department of Engineering and System Sciences at National Tsing Hua University.

余昊翰先生目前為國立清華大學工程與系統科學系博士生。

Hao-Han Yu is currently a Ph.D. student in the Department of Engineering and System Sciences at National Tsing Hua University.

洪明秀女士現為國家衛生研究院生物技術與藥物研究所研究員。

Ming-Shiu Hung received her Ph.D. in Cell and Developmental Biology from Rutgers University, USA. She is currently an Investigator and Associate Director at the Institute of Biotechnology and Pharmaceutical Research, National Health Research Institutes.

曾繁根先生加州大學洛杉磯分校 (UCLA) 機械航空工程博士現為國立清華大學工程與系統科學系講座教授，以及中央研究院應用科學研究中心研究員。

Fan-Gang Tseng received his Ph.D. in Mechanical Engineering from the University of California, Los Angeles (UCLA). He is currently a Chair Professor in the Department of Engineering and System Sciences at National Tsing Hua University and a Research Fellow at the Research Center for Applied Sciences, Academia Sinica.

高性能 ScAlN/GaN HEMT 的製作關鍵

Fabrication Keys for High Performance ScAlN/GaN HEMTs

洪茂峰

Mau-Phon Houg

ScAlN/GaN 因強烈的自發及壓電作用產生的極化效應會在介面形成高密度的 2 DEG ($> 6 \times 10^{13} \text{ cm}^{-2}$)，此特性帶來良好的靜電控制能力與可強化載子侷限效果的優勢，使 ScAlN 元件展現出優於 GaAlN/GaN 的高崩潰電壓及優異的高頻性能，顯示其於下世代高功率與高頻元件應用的潛力。然在評比 ScAlN/GaN 與 GaN-基元件的 RF 表現時卻發現前者在一些特性仍遜於後者，其原因可能是前者尚未建立如後者般成熟的技術，故本文盤點了製作高性能 ScAlN/GaN HEMTs 所需的關鍵技術，並提供可呈現優秀 RF 表現的 GaN-基元件的製作細節，期有助於高性能 ScAlN 元件的製作。惟分析近期 ScAlN/GaN 元件的發展，發現其特性提升的障礙竟然是 ScAlN 本質極化作用的限制，因此本文提出製作高性能 ScAlN/GaN 元件的新思維，即善加利用極化特性並輔以電場重分布技術與發展阻抗工程技術，將其與元件結構和寄生效應等參數耦合，以之建立可於 RF 應用的高性能 ScAlN/GaN 元件，如 LNA 或 PA，等電子系統的製作。

ScAlN/GaN HEMTs exhibit strong polarization effects that induce a high-density 2DEG ($> 6 \times 10^{13} \text{ cm}^{-2}$) at the heterointerface, enabling superior electrostatic control and enhanced carrier confinement. These advantages result in high breakdown voltage and excellent RF performance, demonstrating potential superiority over AlGaIn/GaN devices for next-generation high-power and high-frequency applications. However, comparative studies reveal that ScAlN/GaN devices still underperform GaN-based counterparts in certain RF characteristics, primarily due to the lack of mature technologies. In this work, key techniques for achieving high-performance ScAlN/GaN HEMTs are reviewed, alongside fabrication insights from state-of-the-art GaN-based RF devices. Despite recent progress, critical challenges remain, including relatively low μ and excessive gate leakage. These limitations are associated with the intrinsic polarization properties of ScAlN. Therefore, this study proposes a new strategies in which a co-optimization of polarization engineering, electric field management, and impedance design is required to couple device structure, polarization effects, and parasitic elements. Such an approach enables the realization of high performance ScAlN/GaN devices for RF applications, including low-noise amplifiers (LNAs) and power amplifiers (PAs).

一、高頻高功率的電子系統應用現況

隨著 AI 資料中心、車用電子、工業控制與能源管理系統對高效率電能轉換需求的快速成長，GaN 因為是寬能帶材料且具備更高的功率密度與熱傳導效率，在相同的體積下可提供更大的功率、更低的能耗並使元件小型化，尤其以之製作的側向式 AlGaIn/ GaN HEMT 藉由 field plates back-barriers 與 graded buffers 的設計可將其工作電壓提升達 650 V，並以較 Si 元件更高的電子移動率的優勢令系統可操作於數百 MHz 級頻段進行高效的切換，因此適用於伺服器的 DC-DC converter 或高功率且低損耗的消費性產品等寬廣的產品線，故快速地取代傳統 Si 元件的應用領域。

然而，近年來受惠於 AI 技術的蓬勃發展，AI 資料中心為提升能源效率需將電源架構提升到更高的 800–1.2 kV 範圍，如輝達已規劃下一代的 AI 伺服器將採用 800 V 的高壓直流 (HVDC) bus。另一方面，下世代電動車 (EV) 也為因應更長的續航能力、更快的充電速度以及更大功率輸出的需求，而規劃將 EV 電池架構從 400 V 提升到 800 V。另一方面：目前主導太空產業的推手，包括 SpaceX、歐洲衛星業者 OneWeb，以及亞馬遜的低軌衛星 (LEO) 網路等企業的積極發展也帶動了對高效能 RF 元件的需求。這類晶片必須能處理微波到毫米波的 Ku 或 Ka 頻段且進行高速資料傳輸時須具備寬頻且低延遲失真的特性，同時也要承受太空的嚴苛環境，而寬能隙的 GaN-基 HEMTs 恰具備這種特質，使其可以擔負起下世代高功率與毫米波通訊系統的重要角色備受期待。

二、GaN-基 HEMT 與 SiC 元件的性能比較

雖然 GaN-基 HEMTs 被認為在 RF 與高功率的表現可優於 Si 元件，但預期其可以主宰 EV 動力驅動的逆變器或大型 AI 資料中心的電源轉換器等高功率的應用恐怕是太樂觀了，因為在這個領域還有一個可敬的對手，即 SiC-基的元件。以目前商用功率元件的規格來看，GaN HEMTs 的耐壓能力只約 100–650 V，但 SiC MOS 卻高達 650–3300 V，難怪其可以主宰 EV 的牽引系統。兩者的差距這麼大是有多重的原因的，但最主要的還是元件結構的差異。由於商用的 GaN HEMT 功率元件大都為側向式結構電流沿元件表面流動，所以電壓阻斷是發生在橫向，若要提升耐壓程度則需增加電流漂流的長度，但此舉將犧牲晶片面積與導通效率，因此限制了其耐壓能力的提升。反觀 SiC MOSFET 為垂直式結構，電流流動方向為垂直晶片表面，故電壓阻斷是發生在垂直方向，若要提升耐壓能力只需增加膜厚即可，使其耐壓的表現遠勝於前者。另外，元件可承載電流大小也是高功率元件的重要規格，GaN 因電流受限於通道寬度，故需要較大尺寸的元件方可承載更多的電流，相較之下 SiC 只需調變膜厚即可承擔遠大於前者的電流。再者後者使用了技術相對成熟的 MOS gate，而前者只能製作金屬 Schottky gate 或 MIS gate，因此常造成 V_{th} 飄移，陷阱捕獲效應與電流崩塌 (current collapse) 等可靠性的困擾。綜合上述，目前在高功率的競爭中，SiC 已通過多種驗證標準也在應用領域占有絕對優勢，然而 GaN 則是力爭上游中。

儘管 GaN-基元件在功率的表現尚落後於 SiC，但 III-N 族半導體材料因擁有極化特性會在元件的異質介面處產生高密度的 2DEG，這是 SiC 元件沒有的特點，而且其電子 mobility 也甚大於後者，約為 $1500-2000 \text{ cm}^2/\text{V}\cdot\text{s}$ 與 $900 \text{ cm}^2/\text{V}\cdot\text{s}$ 之比，這些優勢令前者的 g_m (轉導) 甚大於後者，而 g_m 愈大表放大能力愈好，又 RF 特性的 f_T (電流增益截止頻率) 係正比

於 g_m ，顯然 GaN-基 HEMTs 在高頻與低切換損等表現會優於後者，將可在伺服器的功率供應，RF 通訊與電源快充裝置等應用攻佔一席之地。因此可以說：GaN-基元件在下世代電力電子系統中的主戰場應定位在高頻與中級電壓的系統，如資料中心的電源供應，或 LEO 寬頻系統的微波到毫米波頻段，等應用，而非 EV 或再生能源系統的數百 kW 到 MW 的超高功率等級。職是之故，本文後續的討論將就如何提升 GaN-基 HEMTs 在 RF 與微波應用的性能為主。

三、AlGaIn/GaN HEMT 現況與瓶頸

儘管 AlGaIn/GaN HEMT 在目前 RF 的應用上已有不負眾望的表現，然而隨著 6G 以及衛星通訊系統的發展需求頻率愈高，AlGaIn 元件的窘境畢現。因為 AlGaIn/GaN HEMT 的性能仍受限於 GaN 材質的不穩定，容易在表面及介面處生成缺陷，形成大量的汲極漏電流，使其在高頻或高電流切換時會產生汲極延遲 (drain-lag)，電流崩塌 (current collapse) 與嚴重的動態特性退化 (dynamic losses)，除造成工作時切換效率變差同時也造成可靠性的困擾。再者，AlGaIn/GaN HEMT 的閘極技術也是個棘手的問題，因為閘極結構的介質層在經長期重複性的切換工作，會因承受了高應力而產生閘極漏電流與 time-dependent dielectric breakdown(TDDb)，導致閘極失效或 V_{th} 飄移造成元件及系統雪崩式的損壞。這些可靠性的困擾對現行 AlGaIn/GaN HEMT 的表現都造成巨大的壓力也阻礙了 AlGaIn/GaN HEMT 擴大其應用領域的機會。而 AlGaIn/GaN HEMT 既存的表面電場熱點，閘極電場聚集，寄生效應，導致的漏電流與早發性崩潰等困擾都成為其應用於下世代高性能電子系統的障礙。尤其在 RF 實務上，PA 元件 (功率放大器) 的 f_T 與 f_{MAX} (最大震盪頻率) 通常需要高於工作頻率的 3 至 10 倍 (依經驗調整) 才能確保增益、效率與穩定性。對 AlGaIn/GaN HEMT 而言這些要求更是 mission impossible，因此各界須以新材料，新製程或新系統設計的思維來面對這些挑戰⁽¹⁾。

很幸運的，近期針對 ScAlN/GaN 的研究發現其元件特性相對於現今已成熟的 AlGaIn/GaN 有許多的優勢，如超高的極化可以產生高密度的 2DEG，較高的 g_m 使其有更高的 f_T ，並且載子侷限能力很強使它擁有較高的 BV_{DS} (源-汲極間的崩潰電壓)，另外更低的導通電阻 R_{ON} 與導通損耗，減少電路中電感與電容等零件的尺寸與使用的數量可顯著縮小模組的體積，這些令人驚豔的特點對資料中心伺服器，甚至衛星通訊系統的佈局與散熱配置都非常有利，恰可改善現今 AlGaIn/GaN HEMT 所面對的困境，也令 ScAlN/GaN HEMT 成為下世代高頻高功率電子系統的最佳選擇。

四、ScAlN/GaN HEMT 應用於 RF 的潛力

ScAlN 並非新材料其實它來自 AlN，據研究 AlN 為 wurtzite 結構且其 Al-N 鍵較短，若以 Sc^{3+} (離子半徑約 0.75 Å) 取代部份的 Al^{3+} (離子半徑約 0.535 Å) 形成 $Sc_xAl_{1-x}N$ 時，將使其與 GaN 的晶格常數失配 (lattice mismatch) 增大，因此使 ScAlN/GaN 異質介面的壓應變 (compressive strain) 變大，由於壓電特性 (piezoelectric) 與應變成正比且 ScAlN 本身的壓電係數也很高，兩者疊加的結果使介面極化不連續 (polarization discontinuity) 顯著增大，因此在 ScAlN/GaN 介面產生極化電荷 (polarization charges)，並吸引出高面密度的 2DEG (two-dimensional electron gas)，使 HEMT 的導通電阻 (R_{on}) 得以降低、飽和電流可以提升。以下

介紹 ScAlN/GaN HEMT 適於 RF 應用的一些特質⁽²⁾。

1. 相較於一般的 III-N 族材料，ScAlN 擁有特強的鐵電與壓電特性，製成鐵電 FET (FeFETs) 結構的非揮發性記憶體在高頻應用時有優良的表現，若製作 ScAlN/GaN HEMT 時能適當調控能障層的應力 (strain-controlled barrier) 將可增強其功率與 RF 方面的表現。
2. ScAlN 組成為可調，可直接磊晶於 AlGaIn/GaN HEMT 結構而不虞有晶格不匹配或生成介面缺陷的困擾。至於品質良好且完整的磊晶層結構則是元件具備高電子移動率的首要關鍵。
3. ScAlN/GaN HEMTs 因擁有強大的極化特性與優良的介面品質，使其在介面的 GaN 通道中得以產生更多且速率更高的 2DEG，由於電流密度大幅增加，依下式

$$g_m = \frac{\partial I_D}{\partial V_{GS}} \quad (1)$$

ScAlN/GaN HEMTs 將呈現極大的 RF 轉導 g_m 。又因元件的 f_T 與 g_m 呈正相關，這些優勢都使 ScAlN/GaN HEMTs 較 AlGaIn barrier 的元件呈現了更優良的功率，切換速度與效率等表現。

4. ScAlN/GaN HEMTs 的 π PB (功率－頻寬乘積) 約為 GaN 元件的 5－7 倍，也是 SiC 元件的 8－11 倍，由於 π PB 是評估功率放大器在功率輸出與工作頻寬綜合能力的優值指標，數值愈大代表此元件的物理性能上限愈高，而在技術比較或選材評估上通常是取 π PB 愈大愈好，因此預期 ScAlN/GaN HEMTs 高頻表現將優於 AlGaIn 或 SiC，可滿足 6G 與衛星通訊等需求。
5. 另一個重要的概念則是元件的閘極長度 (Lg)：通常 Lg 愈短會令操作頻率愈高，而最先進的 AlGaIn/GaN HEMT 已可製作 50－100 nm 的 Lg，那 ScAlN 又是如何呢？據 P.Cui 的研究顯示⁽³⁾，其團隊將具有強極性的 InAlN/GaN HEMT 的 Lg 微縮到極短的 20 nm 時可將 f_T/f_{MAX} 推進到超高的 320/535 GHz，由於 ScAlN 與 InAlN 的極性表現最為相近，故被認為微縮能力應是相當，預期應會有同等優良的表現。值得一提的是 AlGaIn/GaN HEMT 若微縮其 Lg 至 100 nm 以下時會衍生強烈的短通道效應 (short-channel effects, SCE) 劣化其表現。ScAlN/GaN HEMTs 也會發生這個困擾嗎？據研究 ScAlN HEMTs 不虞會有 SCE 的困擾，因為其 Lg 即使微縮到 20 nm，強烈的極化特性仍令閘極對通道電位有良好的控制能力，甚且再將 ScAlN 能障層薄化到 5－10 nm 時還能更縮短等效的 Lg，使靜電耦合更強，也更強化閘極的控制，所以 RF 表現反而更佳。但必須注意的是：水能載舟也會覆舟，後續的討論會提及 ScAlN 強烈的極化特性在微縮過程會衍生一些寄生效應導致 f_{MAX} 的表現不如預期，所以極化作用的處理也是製作 ScAlN/GaN 元件時不容忽視的關鍵。

五、ScAlN/GaN HEMTs 性能提升的準則

一般而言，應用於 6G 的 sub-THz 通訊系統較重要的特性需有： f_T 要高於 200－300 GHz， f_{MAX} 得大於 400 GHz，功率密度至少要有 10－30 W/mm，功率附加效率 (PAE) 效率則為 50－60% 以上，而且 RF 功率放大器 (PA) 的 BV_{DS} 須達 30－70 V 的水準。這些規格對 AlGaIn/GaN HEMT 目前的技術能力造成不小的壓力，而 ScAlN/GaN HEMTs 因其較 AlGaIn

更佳的材料優勢故被期望能改善這個困境，那要如何著手才能擔起接棒的任務呢？以下整理了近期 GaN-基元件相關的研究策略，有些是 ScAlN/GaN 元件的成果，有些則是 AlGaIn/GaN 等元件的經驗，以之作為 ScAlN HEMT 研發方向的參考。

1. ScAlN/GaN HEMTs 的 BV_{DS} 提升策略

首先來看崩潰電壓， BV_{DS} 這個參數。在 RF 系統中 BV_{DS} 是個很關鍵的特性，因為 BV_{DS} 越大，功率放大器的汲極電壓可以操作的空間會更大，而因輸出功率 P_{sat} 與 V_{DS}^2 呈正相關，當然就使輸出功率隨之增大了。另外，較大的 BV_{DS} 也方便讓系統操作於可匹配最佳的負載阻抗而提升 PAE，增進系統效率。總之， BV_{DS} 越高，元件就能耐越高壓，而轉換出更強的功率，這正是達成 6G 長距離、高覆蓋率 (high EIRP) 的關鍵。由研究可知， BV_{DS} 可藉由適當設計元件的 gate 與 drain 距離 (G-D) 來達成，因為 G-D 會影響電場分佈，若 G-D 距離較短：電場將集中在很小的區域內，導致崩潰電壓較低，反之若加大 G-D 距離電場被分散則使崩潰電壓提升，但其代價就是同時增大導通電阻與元件尺寸，也就是功率元件製作時必然會遭遇 R_{on} 與 BV_{DS} 相互 trade-off 的困擾，因此製作 RF HEMT 時 G-D 須取其短，而 power HEMT 時卻又須取其長。如高功率的 ScAlN/GaN HEMTs 其 BV_{DS} 可達 1 kV 以上，但在 RF HEMTs 的 BV_{DS} 則是小於 100 V，這麼大的差距就是將功率元件的 G-D 間隔設計為 10–40 μm ，而 RF 元件的 G-D 則為 0.5–1.5 μm 即可。顯然在尺寸受限制的條件下製作可以兼具優秀的 RF 與 BV_{DS} 特性的 ScAlN/GaN HEMTs 就成為一大挑戰了。然而 GaN-基元件有望緩解這個 trade-off，因為 GaN 的臨界電場遠高於 Si 且其 2DEG 又具高遷移率與高電荷密度，使具同樣 BV_{DS} 的 GaN-基 HEMT 的 R_{on} 可以較 Si 元件更低 1–2 個數量級，這個優勢增加了 ScAlN HEMT 應用於下世代高性能電子系統的潛力。至於 ScAlN 用以改善這個 trade-off 常用的方法有：採用 field plate 讓 G-D 間的電場更均勻或提升磊晶膜品質以增加 2DEG 電荷密度與遷移率，都能有效提升元件的輸出功率。這些方法詳述如下⁽⁴⁾：

1.1 GaN 緩衝層修飾

由 AlGaIn/GaN/AlGaIn 的經驗，可知其介面量子井會形成具有電子阻斷功能的 EBLs (electron blocking layer) 有效侷限 2DEG，使其免於外溢形成漏電流或早發性崩潰等不良後果。故有研究認為使量子井更深化應有助元件 BV_{DS} 的提升，緣此有研究利用 p-type 摻雜的 GaN buffer 來提升 ScAlN/GaN HEMT 的導帶能障以之強化對介面 2DEG 的侷限，果然有效壓抑了元件的穿透漏電流 (punchthrough leakage)，將 ScAlN/GaN HEMTs 的 BV_{DS} 提升到 640–750 V。另，高品質的 GaN 緩衝層因可消除缺陷引發的崩潰，故也有研究採用了薄層的 GaN 與背障層 (back-barriers) 的組合，除可減少 trap 引致的漏電流外，另背障層也貢獻了侷限通道內的電子減少其外漏並使遷移率提升的功能，果然也將 BV_{DS} 有效提高。

1.2 元件參數與 Field Plates 的設計

有研究提出，藉由適當的元件幾何設計，如閘極金屬的高度，場板 (field plates) 與電極間隔等都可將 GaN-基元件的 BV_{DS} 提升到 kV 等級，如 Wang⁽⁵⁾ 就將 AlGaIn/GaN HEMTs 中間極金屬高度從 0.8 μm 增高到 1.6 μm ，並輔以場板的設計，結果將元件的 BV_{DS} 從 1535 V 提升至 1763 V，他們解釋：這是因為閘極金屬高度改變使端點的電場因空乏區寬度延伸重新分布而變得較為平滑，避免了早發性崩潰的可能，而且元件的 sidewall 電容也會被增大

之故。準此，若將此策略應用於 ScAlN 元件預期也應有效。另有研究則是著手於場板的改良，他們發現不論是採用單層或多層 SiN 的場板，並增大 G-D 間距 (l_{gd}) 以及製作良好的保護披覆層，如 SiN/BZN/SiN 疊層，都能消除電場分布的熱點使之更平滑，也能降低 G-D 間的漏電流使 BV_{DS} 提升。

1.3 最佳化的磊晶

有研究指出以分子束磊晶 (Molecular Beam Epitaxy, MBE) 成長 18% Sc 的 ScAlN/GaN 異質介面，因其為晶格匹配，故使缺陷降低了 80% 也減少了漏電流路徑，並因介面為完整陡峭 ($RMS < 1\text{ nm}$)，使得介面極化誘生的 2DEG 密度可以高達 $6 \times 10^{13}\text{ cm}^{-2}$ ，又因散射減少使移動率上升，HEMT 的輸出電流因而大幅增加，並將 BV_{DS} 提升到 755 V。另還發現，ScAlN/AlGaN HEMT 的能障層厚度為 5–10 nm 時，因 ScAlN 呈現最佳的極化效果，使 HEMT 整體表現也更理想。

1.4 接觸電阻極小化

降低接觸電阻對功率傳輸的損耗也是研發的重點，Yang 的論文⁽⁶⁾指出：他們製作了 L_{gd} 與 L_g 各為 $1.0\text{ }\mu\text{m}$ 與 70 nm 的 AlN/GaN RF HEMT，發現此元件在 V_D 為 5V 時其 f_T/f_{MAX} 高達 140/301 GHz，雖然 BV_{DS} 可能會遭受 RF HEMTs 強烈場聚集 (field crowding) 的干擾，但仍有 40–80 V。分析此成果優異的原因可能是因 regrown 的歐姆接觸會將接觸電阻 R_c ，源/汲極擴散電阻 R_s/R_d 以及界面電阻 R_{int} 等不受閘極控制的電阻分量 access resistance 大幅降低，改善 g_m 所致，使 RF 表現得以提升。又因低電阻也減小了焦耳熱效應使 BV_{DS} 不致於受太大的影響。而且以 regrow 磊晶方式也可降低高 Al 含量的 ScAlN 能障層中產生游離缺陷的機會，避免了 Schottky barrier pinning 的困擾，有利於 ScAlN HEMT 於高功率的表現。

1.5 採用 High-k 保護披覆層

Cheng 研究⁽⁷⁾指出：採用 SiN/BZN/SiN 或是 ScAlN/ScN 等高介電常數 (high-k) 夾心疊層作為 ScAlN/GaN HEMT 元件的閘極鈍化層結構時，由於 BZN 等高介電層可降低閘極與汲極間的峰值電場，減小了 tunneling current，並可壓抑短通道效應與 G-D 漏電流，使 HEMT 元件呈現較佳的能效與耐壓表現。Cheng 還指出其 HEMT 在 L_g 為 200 nm ， L_{GD} 為 $1\text{ }\mu\text{m}$ 時，崩潰電壓從 68.9 V 被提高至 121.5 V，且元件的功率優值 (BFOM) 也達到 $62.6\text{--}87.3\text{ GW/cm}^2$ ，遠高於 GaN-基元件的表現，顯示 High-k 的保護披覆確可有效提升 ScAlN 的高功率表現。

1.6 極化效應與幾何結構的處理

Krause et al. 的論文⁽⁸⁾研究了 ScAlN/GaN HEMT 與 AlGaN/GaN HEMT 兩種元件截止狀態的崩潰現象 (hard off-state breakdown)，發現其 BV_{DS} 分別為 75 V 與 95 V，前者顯然較小。其實兩者的幾何結構如 L_{gd} ， L_g 皆為 $1.2\text{ }\mu\text{m}$ 與 $0.15\text{ }\mu\text{m}$ ，也都有雙場板可將 L_{gd} 附近的電場重分布避免熱點的發生，並設計了 AlN 中介層以強化電子侷限減少漏電流的路徑，更覆上了 SiN_x 保護披覆，但 ScAlN/GaN HEMT 的 BV_{DS} 雖有改善但僅為 75 V，Krause 認為這是因為 ScAlN 的強極化在閘極邊緣 (gate edge) 衍生了強電場，加速其更早發生崩潰之故，這正是水能載舟也可覆舟的寫照。儘管如此，製作 ScAlN/GaN HEMTs 時只要遵循前述準則，如：製造晶格匹配的 ScAlN 能障層，降低導通電阻，設計適當的保護披覆與良好的

閘極工程以處理極化效應等，都可以使 ScAlN/GaN HEMTs 因擁有高密度的 2DEG 與良好的電子移動率，使其能貢獻遠優於 AlGaIn/GaN HEMTs 的 RF 表現。

2. ScAlN/GaN HEMTs 的高頻特性提升策略

一般在研發通訊元件能力時 cutoff frequency (f_T) 與最大 oscillation frequency ($f_{\max}$) 是最被關注的兩個參數，但元件在高頻時會出現輸出功率與頻寬相互 Trade-off 的尷尬情形，故工程師會以功率－頻寬指標 π PB (Power-Bandwidth Figure of Merit) 來評估元件在高頻時維持「高輸出功率」的能力。已知 ScAlN/GaN HEMTs 因特殊的極性電荷效應可提供較大的電流密度，較高的 $f_{\max}$ 與較小的閘極電容，故有利於增大 π PB 值，理論估計其約為 AlGaIn/GaN HEMTs 的 1.45 倍，顯示高頻輸出表現會更優異，也被期待可應用於 6G 與衛星通訊等系統。文獻中有許多提升 ScAlN/GaN HEMTs 高頻特性的方法，其中最被肯定的為：改善磊晶與介面的品質，通道尺寸的微縮與寄生效應的消除等策略，以下分別詳述：

2.1 改善磊晶與介面的品質

如前述，磊晶與介面的品質的改善除有利於提升元件的耐壓特性外，對 HEMTs 的 RF 特性助益尤大。如有研究提出：在 ScAlN/GaN 介面增加一薄層，約 2 nm 的 AlN 中介層，可改善介面品質減少散射的效應，使電子移動率由 $300 \text{ cm}^2/\text{V}\cdot\text{s}$ 大幅提升至 $1573 \text{ cm}^2/\text{V}\cdot\text{s}$ ，有效改善了 g_m 使 f_T 得以提升。另有研究提出：MOCVD 長晶時可使用較高的溫度以減低 carbon/oxygen 雜質造成的缺陷，降低 2DEG 被散射的機會，藉由 2DEG 移動率的改善來提升 $f_T/f_{\max}$ 特性。

2.2 微縮閘極尺寸及降低接觸電阻

如前述 (1.4) 所述，降低接觸電阻確可改善 HEMTs 的耐壓能力，但 Cornell 團隊的研究顯示，若以微縮閘極尺寸及降低接觸電阻並用的策略也有助於提升 RF 特性。如他們就利用次微米的閘極並以 regrown 重摻的 n^{++} -GaIn 降低 S/D 的歐姆接觸電阻，而將 $f_T/f_{\max}$ 大幅提升並使元件在 GHz 頻段操作的 PAE 呈現更佳表現^(9, 10)。

2.3 設計適當的能障層以消除應力

前述 Yang 的研究曾提出⁽⁶⁾：製作元件時將 Sc 含量控制在晶格匹配的 $\sim 15-18\%$ ，薄化 ScAlN 能障層至 3 nm，以 AlN 中介層並披覆具壓應力特性的 SiN_x 保護層以提升導帶量子井對 2DEG 的侷限能力，藉由這些策略可改善 ScAlN HEMT 的高功率表現，其實也有助於提升 RF 的表現，因為這些措施可壓抑 SCE，強化 g_m 的線性度，並將 $f_T/f_{\max}$ 提升至 140/301 GHz 的高水準。模擬甚至顯示若再增加背障層，如 AlInGaIn 與製作 local doping，以降低閘極寄生電容 (C_{gg})，甚至可將 $f_T/f_{\max}$ 推向更高的 300 GHz 以上。

2.4 基板與熱效應的處理

有研究指出：ScAlN/GaN HEMT 若使用高熱導性的基板，如 SiC，並輔以高散熱能力的介電層，如鑽石薄膜，將可維持元件工作於高頻而不虞發生自我加熱的現象，如 Nomoto 就報告了此元件可以獲得大於 4 A/mm 的輸出電流，且其 $f_T/f_{\max}$ 也高達 92.4/134.3 GHz，顯著改善元件的 RF 特性。

此外，另有報告也顯示了改善的成效，如 Chumbes⁽²¹⁾ 只單純降低了 AlGaIn/GaN 電晶體的接觸電阻 ($\sim 0.1 \Omega \cdot \text{mm}$)，就使元件傳送的功率大幅增加到 3.6 W/mm 以上，且在 94 GHz 時的 PAE 可達 24%，其表現較傳統的 AlGaIn/GaN 毫米波放大器更為優異。而 Krause et al. 則藉由改善磊晶品質以提升元件表現，更將可操作於 30 GHz 頻段的 ScAlN/GaN HEMTs 的傳送功率與效率推向 8.4 W/mm 與 48% PAE 之高，且 π PB 更高達 176.5 GHz·W/mm 的水準。綜合來看，可呈現最先進成果的 ScAlN/GaN HEMTs 其製作大都不脫上述的準則，如閘極工程，歐姆接觸與保護披覆製作等技術改善了元件的電子傳輸，功耗與寄生效應所致。

ScAlN/GaN HEMTs 與其他 GaN-基 HEMTs 的 RF 表現比較

如上述已知 ScAlN/GaN HEMT 具有優越的結構特性若再輔以優化的製程，預期表現應可超越傳統 GaN-基 HEMTs，但實情是這樣嗎？表 1 整理了近期相關元件典型 RF 特性研究的成果^(6, 9-12)。從表中可發現，ScAlN /GaN HEMT 因強烈的極化特性使其在導通電流以及輸出功率的表現較佳。但 g_m 卻無法承襲其擁有較大 I_D 的優勢而襄助增大電流增益，如 (式 2) 所示，表現與後者差異不大。

$$f_T = \frac{g_m}{2\pi(C_{gs} + C_{gd})} \quad (2)$$

表 1. ScAlN/GaN HEMTs 與其他 GaN-基 HEMTs 典型 RF 表現比較。

RF 特性	ScAlN/GaN HEMT	GaN-基 HEMTs
$I_{d,max}$ (A/mm)	4.07 A/mm ($L_g = 90 \text{ nm}$, $L_{sd} = 600 \text{ nm}$)	$\sim 1.7\text{--}2.8 \text{ A/mm}$ (AlGaIn/InAlN HEMT) $\sim 2.6\text{--}2.8 \text{ A/mm}$ (InAlN/InAlGaIn HEMT)
g_m (S/mm)	0.50 S/mm ($V_{ds} = 5 \text{ V}$)	0.36–0.66 S/mm (AlGaIn/InAlN HEMT)
f_T / f_{MAX}	92.4 GHz / 134.3 GHz	f_T 達 250–450 GHz (AlGaIn/InAlN HEMTs) f_{MAX} 大於 300 GHz (最佳化設計 AlN/GaN)
BV	BV 為 78 V (當 $L_{gd} = 3.85 \mu\text{m}$) ; BV = 35.5 V (當 $L_{gd} \approx 455 \text{ nm}$) 其崩潰電場約為 0.2–0.78 MV/cm	最佳的 AlGaIn/AlN HEMTs 的崩潰電場可高於 2MV/cm，使其較同等 L_{gd} 的 ScAlN 元件擁有更大的 BV。

影響所及甚至使 RF 速度與耐壓能力落後後者。為何 ScAlN 不能憑藉其強烈的極化優勢而在 RF 的表現上全面領先呢？以下分析幾例具優異成果的 GaN-基 HEMT 研究探討其製作的關鍵：

1. 2019 年 A. Hickman 團隊發表了 AlN/GaN/AlN HEMTs 高崩潰電壓的成果，此元件的 L_{GD} 為 $04 \mu\text{m}$ 且無設計可分散電場的場板，但 BV_{DS} 卻可達 78 V，估計其 G-D 間的等效電場約 2 MV/cm。比之文獻中約當尺寸 ScAlN/GaN HEMTs，的 BV_{DS} 僅為 20–50 V，G-D 間等效電場也為較低的 1 MV/cm，顯見 Hickman 的元件耐高壓能力確遠高於他者⁽¹²⁾。探究 A. Hickman 的製程可發現幾項重要的關鍵，如 (1) Hickman 以 MBE 成長品質良好的

AlN、GaN 等磊晶膜，有效壓抑閘極與汲極導致的漏電流。(2) HEMT 結構中使用了不含 Sc 的 AlN 能障層免除了應力的困擾，也不會滋生 ScAlN 最被人詬病的「極化緩解」或由壓電相變成鐵電相的「鐵電不穩定」的可靠性問題。(3) AlN 生成的 2DEG 不像 ScAlN 那麼高，故電場分布較均勻不會聚集成熱點，降低了早發性崩潰的機會也有助 BV_{DS} 的提升。(4) S/D 區為 MBE regrowth 的 n^{++} -GaN 並以 Non-alloyed Ti/Au 作為金屬接觸，大幅降低了接觸電阻值。這些製程的改善可說是 BV_{DS} 提升的關鍵也不脫前述的準則。但另有一重點值得一提，即 Hickman et al，沒有製作 gate sidewall 的保護披覆，這一點被認為是 BV_{DS} 提升的關鍵，因為 gate sidewall 的損傷被認為是漏電流最主要的來源，也是 BV_{DS} 的最大殺手。而 Hickman 則透過元件結構設計與材料選擇的策略來消除這個難題。如：他們將閘極金屬直接鍍在 AlN 薄層上，如此就無須再以 ICP 作 gate recess 蝕刻，避免了 gate sidewalls 因高能電漿離子撞擊而受損傷。另又採用 AlN/GaN/AlN 量子井通道結構，使 GaN 通道電場垂直於 buffer 與 AlN，故電場不會在 Gate edge 或 sidewall 聚集避免了早發性崩潰的發生。但美中不足的是，gate edge 卻因極化不連續令電場較強，使耐壓能力降低。至於採用 AlN 也是一著好棋，因為其能隙值較 ScAlN 為大可承受近 2 MV/cm 的側向電場，又可以減少熱電子注入 sidewall 的機會，對 BV_{DS} 的提升大有幫助。綜合來說，Hickman 是以電場處理的方式取代困難的 sidewall 修護製程，使 BV_{DS} 提升，相較之下，ScAlN 能隙值較小，ScAlN/GaN 結構又需進行 Gate recess 蝕刻，但因 Sidewall 具高度敏感性常造成元件特性劣化的副作用，使得 ScAlN/GaN HEMT 典型的 RF 表現遜於 Hickman 的元件，顯然對於 ScAlN 材料本質的問題處理，製程設計恐需有新的思維。

2. 在提升 GaN HEMT RF 特性的研究，L. Li et al 展示了一項在 Si 基板上製作高性能 GaN HEMTs 的技術⁽⁹⁾。他們以 MOCVD 製作 InAlN/AlN/GaN 的 HEMTs 結構，因磊晶層品質良好使其 2DEG 高達 $2.27 \times 10^{13} \text{ cm}^{-2}$ ，電子移動率也是優異的 $1430 \text{ cm}^2/\text{V}\cdot\text{s}$ 。Li 又將元件閘極長度 (Lg) 微縮至 55 nm，S-D 間隔 (L_{SD}) 為 175 nm，歐姆接觸以 MBE regrown 的 Si 摻雜 n^{++} GaN 製作，閘極則是 T 形的 Ni/Au 金屬閘，其元件結構及各相關參數詳如 (圖 1) 所示。此元件的 DC/RF 特性均甚優異，如 f_T/f_{MAX} 為 250 / 204 GHz，峰值轉導 $g_m = 0.66 \text{ S/mm}$ ，又 $I_{D,MAX} = 2.8 \text{ A/mm}$ ，且 $R_{ON} = 0.6 \Omega\cdot\text{mm}$ ，雖然其 RF 表現仍略遜於 SiC 元件，但已遠勝現有低於 150 GHz 的商用 GaN on Si 元件。據分析其關鍵為：(1) Li 以 MBE regrown 的 n^{++} GaN 製作 S/D 的歐姆接觸，將 Rc 大幅降至 $0.07 \Omega\text{mm}$ ，也使 2DEG 通道的 R_{SH} 減小為 $206 \Omega/\square$ ，有效降低了 gate 附近的 access resistance，並因該區域仍擁有相當濃度的 2DEG，得以保有較大的電流增益 g_m ，由 (式 2) 可看到 f_T 被大幅提升至 250 GHz 應是必然，因為 Rc 與 access resistance 的大幅減小，縮短了元件充電所需要的時間，故成為 f_T 提升的關鍵。(2) Li 以 EBL 製作 Ni (50 nm)/Au (240 nm) 的 T-gates；其為長 150 nm，頭部高 290 nm，寬 300–240 nm 的 T 形狀，由於 gates 長度縮短可減小閘極電阻 R_g 與閘極電容 C_{gs} ，此策略改善了功率增益，也將 f_{MAX} 提升至 204 GHz，這個結果對毫米波 PA 的應用非常重要。具估計此元件的 $(f_T \cdot f_{MAX})^{1/2} \approx 226 \text{ GHz}$ 應是目前 GaN-on-Si RF 元件中最佳的成果。顯然，超低的 regrown 歐姆接觸及不使用退火處理故無 alloy spiking 困擾的策略，使 2DEG 得以保持高密度也降低了 Rsh 等寄生效應，這個製程的突破應是 f_T 與 f_{MAX} 提升的關鍵，另微縮元件尺寸與 T-gates 結構，因減少了電子傳輸時間增強了高頻增益，也幫助了 RF 表現的提升。

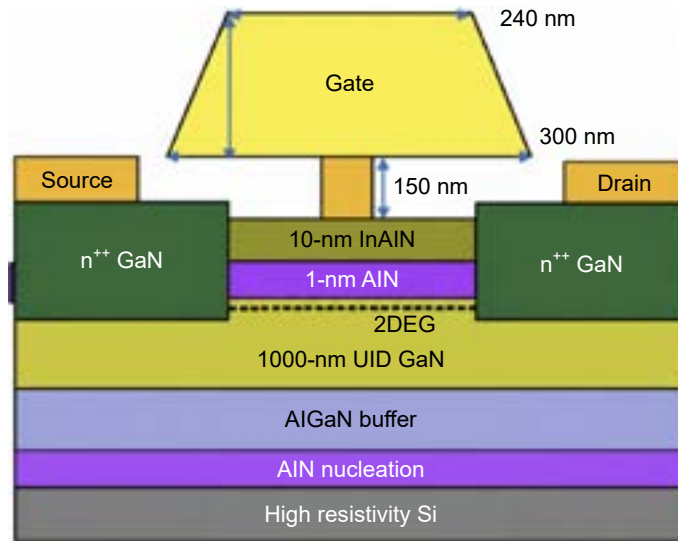


圖 1. InAlN/GaN HEMT 結構，本元件無使用 back barrier。⁽⁹⁾

惟此元件有一美中不足之處，就是其 f_{MAX}/f_T 比值僅 0.82，小於 1。一般而言元件若 f_{MAX} 大於 f_T ，即比值大於 1，表示通道電子的速度可完全轉換為有用的 RF 功率，這才具實用價值。而今 Li 元件的 f_{MAX}/f_T 小於 1，表示元件功率有耗損。據 G. Dambrine 所推導 FET 小信號等效電路中 f_{MAX} 與 f_T 的關係⁽¹³⁾，如 (式 3)，可知損耗的源頭應是來自寄生電阻 R_s 與 R_D 以及 G-D 間電容 C_{gd} 的影響，使 f_{MAX} 無法得利於本質載子的高傳播速度 f_T 而更為提升。顯然 Li 的 RF 元件仍需致力於元件寄生效應的改善。但整體而言，Li 的 GaN-基 HEMT 已有甚高的 f_T 與適當的 f_{MAX} ，且 gate 控制能力佳，雜訊參數也低，應適合 LNA 的應用。惟若要應用於功率放大器 (PA) 則仍需改善寄生效應以提升 f_{MAX} ，這將是其最主要的挑戰。Li 另由模擬發現，若無 regrown 歐姆接觸之助，則此元件的 f_T/f_{MAX} 將大降為 202/171 GHz，顯然接觸電阻的改善對 RF 特性的提升至關重要。另一應關注的重點是，此元件的 BV_{DS} 僅為 12 V，顯然元件在微縮尺寸時也不應忽略 SCE 的影響。

$$f_{MAX} \approx \frac{f_T}{\sqrt{R_G(R_S + R_D)C_{gd}}} \quad (3)$$

3. 一般在評估高頻元件特性時常著重於 f_{MAX} 與 f_T 這兩個參數的表現，而 f_T 通常表示通道電流對閘極電壓反應有多快，是元件本質速度的一個指標，至於 f_{MAX} 則為考慮所有寄生效應的損耗後元件可有效傳送功率增益的最高頻率，顯然它對外在寄生效應是很敏感的。由 (式 3) 可知 R_G 、 R_S 、 R_D 、 C_{gd} 都會影響 f_{MAX} ，而這些寄生效應都與製程相關，故優化製程應有助於其特性的提升。儘管前述 Li 以 MBE regrown 的 Si 摻雜 n^{++} GaN 作為 S/D 區的歐姆接觸改善了 R_C ，使高頻特性顯著提升，但其 f_{MAX}/f_T 比值為小於 1，顯示寄生效應仍嚴重影響著高頻響應。Yang 對此議題提出了解決的方法：他認為在 MBE 磊晶時以高濃度的 Si 進行摻雜時會衍生應力的麻煩，造成漏電流以致無法使 access resistance 更為降低，但如改為 Ge 的摻雜，因其活化程度較 Si 為高故能得到更高的電子濃度，且 Ge 在

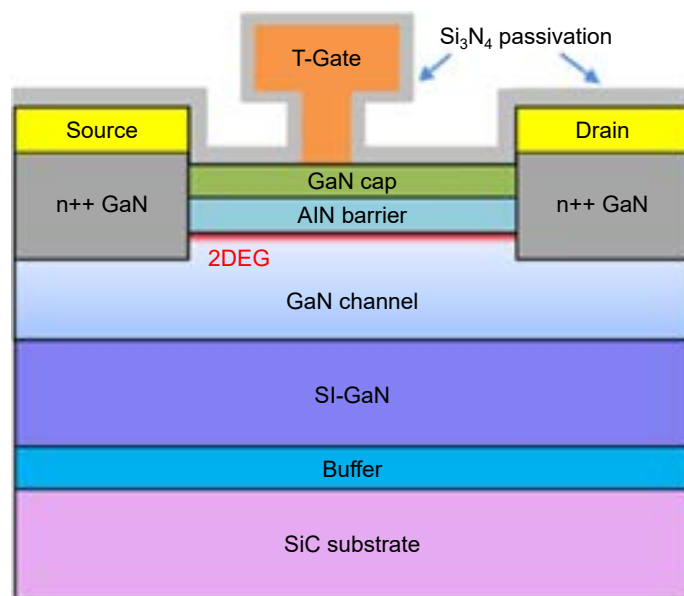


圖 2. AlN/GaN HEMT 結構，以 Ge-摻雜 n^{++} GaN 作為歐姆接觸。⁽⁶⁾

近 regrowth 的介面因應力較小故表面較平滑，缺陷也較少使電子傳輸會更理想，對降低 access resistance 更有效，因此他們在 SiC 上製作 AlN/GaN HEMTs 時遂改以重摻雜 Ge 的 GaN 為 S/D 的歐姆接觸，並將 AlN 能障層薄化至 3 nm 以強化對 2DEG 的侷限，以此策略將元件的 f_{MAX} 推進到超過 300 GHz 的水準，也得到比值大於 1 的 f_{MAX}/f_T 。他們的研究顯示了壓抑寄生效應對 RF 特性改善的重要。Yang 的元件結構 (圖 2) 與關鍵製程概述如下：AlN/GaN HEMT 製作時其 S/D 是以 MBE regrown 重摻雜 Ge 的 GaN 為歐姆接觸，由於磊晶膜品質良好使電子濃度達 $2.5 \times 10^{20} \text{ cm}^{-3}$ ，故將接觸電阻 R_c 降低至 $0.09 \Omega \cdot \text{mm}$ 。量測此元件在閘極長 70 nm 時的 RF 響應，發現其 $I_{D,max}$ 為 1.54 A/mm，峰值轉導 g_m 約 0.74 S/mm， f_T/f_{MAX} 則達 140 / 301 GHz，且 off-state 的 BV_{DS} 亦為 80–120 V 之高水準表現。Yang 的元件有一個重要的成果，即 f_{MAX}/f_T 比為 2.15，這個 f_{MAX} 甚大於 f_T 的結果表示：即使元件操作於 f_T 以上時仍可傳送可用的增益，也就是說此元件不但可操作於毫米波段還可有效傳送 RF 功率，此一成果顯示了 AlN/GaN HEMT 作為毫米波功率 PA 的潛力。

分析 Yang 優異的 RF 成果可歸因於幾個因素：(1) 超高的 Ge 摻雜，高達 10^{20} cm^{-3} ，使歐姆接點具強烈的 tunneling 特質大幅降低 R_c 。(2) 薄化的 AlN 能障層有極強的靜電控制能力，除抑制了 drain-induced barrier lowering (DIBL) 避免 SCE 效應外，也能有效侷限 2DEG 使輸出電導 (output conductance) g_{ds} 減小 (即增大 R_{ds})，使輸出功率增大，這個策略的效果則以 f_{MAX} 的提升表現出來。(3) 在 HEMT 元件的特性中， g_m 可表為 (通道控制能力) / (漏電流) 之比，當 gate 控制通道的能力越好或汲極漏電流愈小都會令 g_m 愈高，但汲極漏電流與 g_{ds} 相關，將之代入 (式 3) 可以導出 f_{MAX} 正比於 $(g_m/g_{ds})^{1/2}$ 的關係。由前述 (1) 與 (2) 的分析，Yang 就是採用了超高的 Ge 摻雜將 g_m 增大再以薄化的 AlN 能障層來降低 g_{ds} ，在這雙重利基加持下果然將 f_{MAX} 推升到 301 GHz 的高水準。Yang 的研究提供了一個重要的概念：即設計 f_{MAX} 的提升時應著重改善 g_m 以及增強通道的控制能力以壓抑 g_{ds} 的影響。

前述介紹的 GaN-基 HEMTs 研究，可發現在不同能障層結構下，如 AlGaN、InAlN、AlN 等，只要適當設計製程都可改善其 RF 特性，然而這些材質各存在一些寄生效應限制著 GaN-基 HEMTs 提升特性的努力。惟若以 ScAlN 材料代之，藉其優異的極化特性能否產生

突破性的成果呢？以下介紹兩個 ScAlN HEMTs 的研究作為比較。

4. 2025 年，Nomoto et al. 成功的製作了可操作於毫米波頻帶並呈現極高電流密度的 AlScN/GaN HEMT 元件⁽¹⁰⁾。他們在 SiC 基板上以 PA-MBE 磊晶成長 2 nm 的 GaN cap，5 nm 的 $\text{Al}_{0.86}\text{Sc}_{0.14}\text{N}$ 能障層，2 nm 的 AlN spacer，1 μm 的 UID GaN 通道與 AlN 成核層 / SiC 等結構，得到 2DEG 為 $3 \times 10^{13} \text{ cm}^{-2}$ ，且電子移動率 $494 \text{ cm}^2/\text{V}\cdot\text{s}$ ，顯示磊晶膜品質良好。接著製作 MBE regrown n^+ GaN 的 S/D (Si: $7 \times 10^{19} \text{ cm}^{-3}$)，將 Rc 降低至 $0.31 \Omega\cdot\text{mm}$ ，再經由 recess 製作了 90 nm 的超短 T-gate，再定義 Lsd 為 00 nm，並藉由縮短 access 區降低 R_{on} 以提升 Id，其元件結構如 (圖 3) 所示。這些精心設計的製程果然使該元件呈現良好的 DC/ RF 特性，如： $I_{\text{d,max}}$ 呈現了目前 AlScN HEMT 的最高紀錄， 4.07 A/mm ，另 R_{on} 為 $0.59 \Omega\cdot\text{mm}$ ， g_{m} 為 0.5 S/mm ， $f_{\text{T}}/f_{\text{MAX}}$ 則為 $92.4 / 134.3 \text{ GHz}$ ，且 BV_{DS} 也是良好的 $35-78 \text{ V}$ 。其中 f_{T} 顯示元件速度已可滿足毫米波頻帶且 RF 特性也不遜於前述 GaN-基元件。Nomoto 分析其原因為：(1) $\text{Sc}_{0.14}\text{Al}_{0.86}\text{N}$ 的組成與 GaN 晶格匹配介面缺陷少，且 $\text{Sc}_{0.14}\text{Al}_{0.86}\text{N}$ 的自發與壓電極化特性較 AlGaN 更大故可產生較高的 2DEG，雖然因散射作用限縮了 mobility，但仍產生高達 4 A/mm 的輸出電流，顯然 ScAlN 強烈的極化特性使元件呈現高電荷濃度驅動型的 HEMT。(2) 傳統 ScAlN 微縮 Lg 時須同時將厚度減薄，然而這卻令 gate 產生大量的漏電流，反觀 ScAlN 為高 κ 介質， $\kappa \approx 15$ ，即使將 $\text{Sc}_{0.14}\text{Al}_{0.86}\text{N}$ 薄化至 5 nm，仍可維持高電容有利於增加 C_{gs} ，且可支撐 90 nm 的閘極，這也是 $\text{Sc}_{0.14}\text{Al}_{0.86}\text{N}$ 材料的優勢使它得以操作於甚高 f_{T} 的原因。(3) Regrown 的歐姆接觸降低了 Rc，且縮短的 Lsd 也令 R_{on} 減小，這是 Id 甚高的主因。(4) 2 nm 的 AlN spacer 讓電子波函數遠離粗糙的界面，由於散射減少可將 g_{m} 維持在良好的 0.5 S/mm 。(5) T-gate 降低了 Rg 且縮短的 Lsd 也降低了 Rs，這些策略都有助於使 f_{MAX} 達到 134 GHz 且令 $f_{\text{MAX}} > f_{\text{T}}$ 。Nomoto 的成果首次證明了 AlScN 極化材料可以製作成兼具「超高電流密度」與「毫米波速度」的 RF 元件，也宣告了 ScAlN

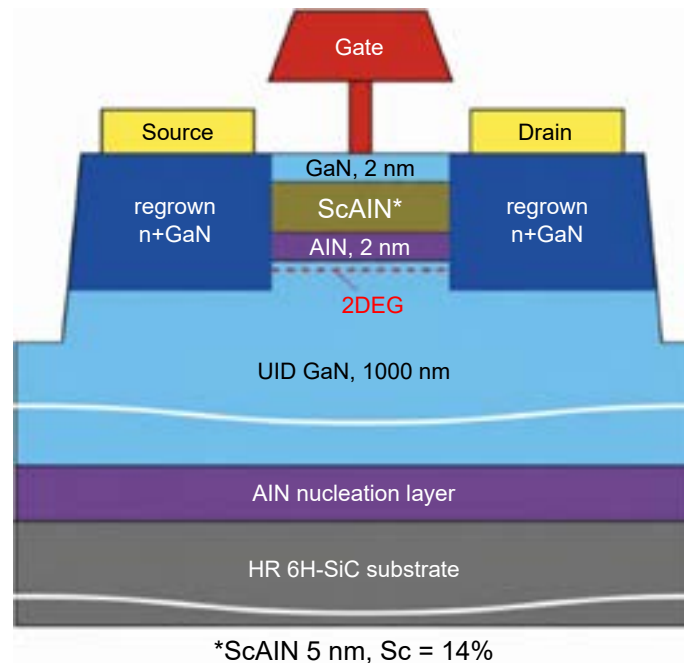


圖 3. ScAlN/GaN HEMT 結構，此元件的能障厚為 5 nm，並以 regrown 製作歐姆觸。⁽¹⁰⁾

HEMT 有潛力可取代 AlGaIn HEMT 實現 6G 與毫米波通訊的應用。

5. 儘管前述 Nomoto et al. 是利用 ScAlN 優勢的極化特性製作了良好表現的 RF 元件，但 Nguyen et al. 卻是以不同的策略，其不求取最大化的極化電荷，而是以良好的磊晶 (strain-balanced epitaxy)，微縮的 T-gates 與消除寄生效應的方法將 ScAlN/GaN HEMTs 由電流增益模式轉變為功率增益模式，使其達到 f_{MAX} 超過 f_T 的表現，此研究發表在 2024 的 IEDM 會議中⁽¹¹⁾。Nguyen 在報告中說明：所謂的 strain-balanced epitaxy 係以磊晶成長一種包含有 5 nm 的 $\text{Sc}_{0.13}\text{Al}_{0.87}\text{N}$ (其生成壓縮式應力) 的能障層與 1.5 nm 的 AlN interlayer (其生成伸張式應力) 的複合結構，此層磊晶膜因內部應力被平衡了故可消除磊晶時產生會損及電子運動或形成漏電路徑的缺陷，而且可確保 ScAlN 能障層保持有高度的極化，得以保有 $\sim 2.5 \times 10^{13} \text{ cm}^{-2}$ 的高密度 2DEG，這個高品質磊晶膜就是使 ScAlN/GaN HEMTs 元件達成高 RF 性能最重要的基礎。Nguyen 接著以 MBE regrown n^+ -GaN 製作 S/D 以降低 R_c 至 $0.09 \Omega \cdot \text{mm}$ ，又因無熱退火製程故減低了介面的粗糙程度。接著製作 40 nm 的 Ni/Au 金屬 T-gate，並以 ICP 蝕刻 GaN cap 與 AlScN 能障層以進行歐姆接觸區的 regrowth，由於 Regrowth 是在蝕刻區內進行，故避免了 ScAlN gate sidewall 受高能 ICP 離子損傷的困擾。其元件如 (圖 4) 所示。經量測元件特性結果為：當 L_g 從 800 nm 微縮至 40 nm 時， I_{max} 也從 1.9 A/mm 升高到 2.8 A/mm，且 g_m 也自 0.37 增加為 0.55 S/mm。甚至在 L_g 為 40 nm 時 f_{MAX}/f_T 更是達創紀錄的 321/173 GHz，比值為 1.85，表現在 f_T 達飽和時仍有功率持續供應 f_{MAX} 使其往上提升，如圖 4(b) 所示。Nguyen 利用小信號模型萃取參數來分析此一表現，發現此時元件的 $g_{m,\text{ext}}$ 為 0.4 S/mm， C_{gd} 與 R_{ds} 也各為超低的 8 fF/ μm 與 247 Ω 。據 Miller effect 指出 C_{gd} 會將輸出端汲極的信號耦合回輸入端的閘極，此一反向回饋作用會降低 f_{MAX} 以及功率增益，但 Nguyen 藉由微縮閘極的長度與 S/D 的間距並使用精密的微影技術製作 T-gate 將 G/D 重疊區縮小，大幅降低了 C_{gd} 的效應，減除了 Miller effect 的困擾，使元件的功率增益在 f_T 飽和之外的頻率仍可持續產生，呈現 $f_{\text{MAX}} > f_T$ 的表現，也確保此 ScAlN HEMTs 的毫米波 PA 可提供穩定的功率增益。

Nguyen 的 ScAlN/GaN HEMT 優異的 RF 特性可歸因於下列的製程：(1) 低 R_c 的 regrown 接觸技術：以 MBE regrown 的 n^+ -GaN 做為 S/D 的接觸，將 R_c 大幅降低至 $0.09 \Omega \cdot \text{mm}$ ，減低了導通電阻與 access 區的損耗使 f_{MAX} 可以提升，這是最關鍵的步驟。(2) 製作超短的 T-gate：以 EBL 將 T-gate 微縮至 40 nm，因 R_g 、 C_{gd} 等寄生效應被減小也使 f_{MAX} 得以提升。(3) 成長應力平衡結構將磊晶層的品質提升：成長應力平衡的 AlN 與 ScAlN 的複合結構，得到表面平滑且品質良好的磊晶膜又降低 ScAlN 能障層的厚度，使 2DEG 維持適當且移動率良好，除方便後續介質保護層的製作外，也有利提升 ScAlN HEMTs 的 RF 性能。

六、高性能 ScAlN/GaN HEMTs 的製作關鍵

由前述的探討可發現：ScAlN/GaN HEMT 的 RF 表現相較於 GaN-基 HEMT 無法呈現絕對的優勢竟然是因為 ScAlN 具有強大的極化場之故，因為極化作用雖然可產生高密度的 2DEG，但同時也增大了 C_{gs} 與 C_{gd} ，不利於提升 f_T 與 f_{MAX} 的表現。再者，高 Sc 含量的 ScAlN 磊晶膜長晶較難，品質不佳缺陷又多，甚至因介面極化太強會導致 band bending，使歐姆接觸難以優化。總之，ScAlN/GaN HEMT 是一種綜合了電性優勢和高頻劣化機制的材料系統，若欲強化其 RF 特性必須從材料源頭到製程全面考量才能畢其功。基於前人的經驗，以下幾點被認為是長晶的關鍵：

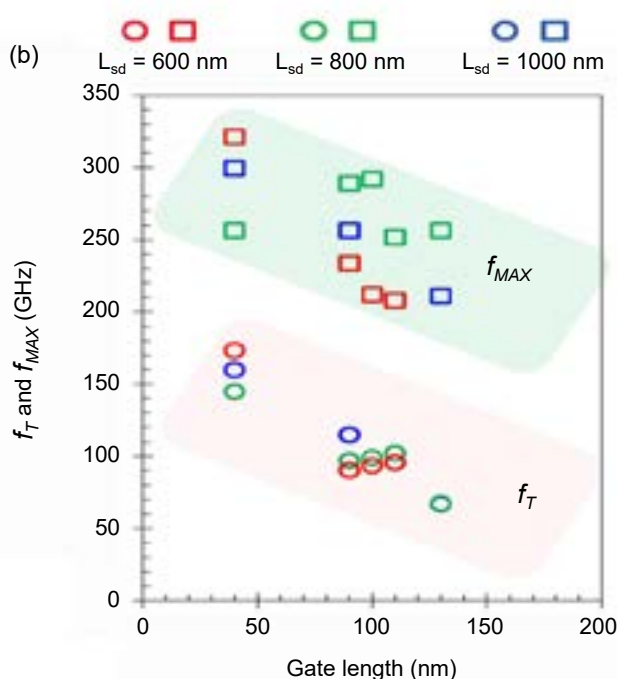
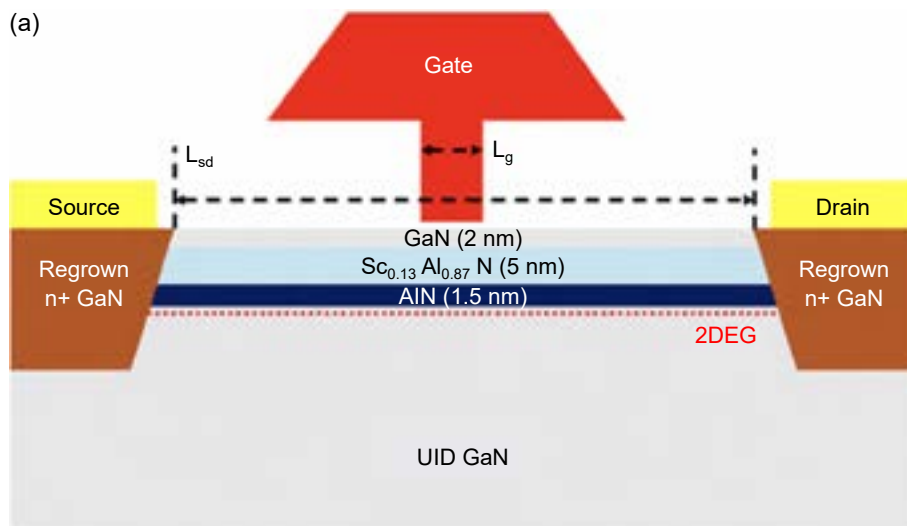


圖 4. (a) Strain-balanced ScAlN HEMT 的結構。(b) f_T 與 f_{MAX} 隨 L_g 的變化，可看到此元件在 T-gate 的 L_g 在小於 130 nm。時呈現 f_T 大於 80 GHz 且 f_{MAX} 高達 210 GHz 以上。⁽¹¹⁾

1. 高品質的 ScAlN 磊晶膜

目前 GaN-基 HEMT 的結構大都為 AlGaIn/AlN/GaN，其特性表現佳且技術已成熟，若欲以 ScAlN 取代 AlGaIn 擔任 HEMT 能障層的角色則 HEMT 的特性會有怎樣的變化呢？T. Okuda 對此作了詳細的探討⁽¹⁴⁾。他發現 ScAlN 層中 Sc 含量增加會改變 wurtzite 的結構，破壞磊晶膜結構的完整性，使 ScAlN 表面變為粗超。另外極化效應雖使 ScAlN 產生高密度的 2DEG 但因散射機率增大反而令移動率 (μ) 降低，劣化電性。進一步研究元件的表現，Okuda 發現，傳統 AlGaIn/GaN HEMTs 的 2DEG 雖較低，但因 μ 較高，故 RF 表現較好，反

觀成長了 ScAlN 能障的 HEMT 雖然有較高的 2DEG，但 RF 表現反而遜於前者，Okuda 認為原因是：ScAlN 極化的紅利被磊晶膜中的缺陷遮蔽了，使 μ 劣化且元件特性也變差。因此 Okuda 認為 ScAlN 磊晶品質對 HEMT 特性有絕對的影響，另外 HEMT 特性與 ScAlN 能障層相關的要求還有厚度，粗糙度，漏電流， μ 等多項，這些議題都與 AlScN 中 Sc 的含量有關，因此成長高品質的 ScAlN 磊晶膜的關鍵顯然就在 Sc 原子的處理。但 Sc 的處理有那些困難呢？

話說 ScAlN 的結晶相為 wurtzite，在熱力學上這是一種介穩態 (metastable)，不穩定又容易產生相分離，尤其在高 Sc 組成，如 Sc 大於 20–25% 時，結晶相會分離成 rocksalt ScN 與 wurtzite AlN 的複合結構，有時則是在 wurtzite 結晶中出現含 cubic 的內容物，這種結晶會影響 ScAlN 元件的最大極化能力，並生成大量漏電流且造成 trap-assisted tunneling 與 V_{th} dispersion 並影響 RF noise 等重要特性。甚且 Sc 的併入還會導致高密度點缺陷生成，使超薄 ScAlN 能障層 (小於 6 nm) 的 HEMT 會在 RF 的高電場操作時生成大量的 gate 漏電流並使元件的 BV_{DS} 降低，如前述 Nomoto et al. 的元件會受 G-D 間漏電流的影響使 BV_{DS} 不佳就是這個原因。因此目前 ScAlN 元件的製作大都建議以 10–18% 為限，取其可以有較佳結晶品質之故。因此成長高品質 ScAlN 磊晶膜的關鍵就在 Sc 原子的處理。下列摘錄了處理 Sc 原子的重點⁽¹⁵⁻¹⁷⁾：

- (1a) Sc adatoms (adsorbed atom) 對 ScAlN 膜的影響：在 ScAlN 成長時，若 Sc 原子過多形成 metal-rich 面時會使 Sc 傾向偏析而造成相分離，並因晶格受扭曲而產生缺陷與粗糙的表面，故一般都採 N-rich 的成長條件或以 pulsed MBE 法來激發 Sc adatoms 的活性，壓抑其偏析，使其能快速生成純相以改善結晶的品質。
- (1b) 精準控制 Sc adatoms：適當調控 V/III 比，即 N/(Sc + Al) 的流量比：如 (1a) 所述，過多的 Sc 會導致生成二次相劣化晶體品質。對此 Pulsed MBE 可以精準調控表面上 adatom 的數量，使其成長高品質的 ScAlN 晶體。另成長溫度與 原子流量的平衡也會影響磊晶表面的動力學反應與 adatom 的擴散，對於晶體維持 wurtzite 相的穩定與均勻甚為重要。惟要特別注意的是：控制 Sc adatoms 時仍須維持 N 原子流量的穩定以利 Sc 併入介穩態的 wurtzite 相。
- (1c) V/III 比為成長 ScAlN 晶體的重要參數：V/III 比，即 N / (Sc + Al)，增加可改善晶體品質使缺陷減少，反之，低 V/III 比，即 metal-rich 將造成粗糙的表面與相分離使晶體品質劣化。但最佳的 V/III 比卻與基板溫度與磊晶方式有關，原則上務必取得足夠的氮成分以穩定 wurtzite 的結晶，並允許適量的 Sc 與 Al 併入 ScAlN 晶體而仍不會造成晶相偏析。總之，精準控制 V/III 比是成長高品質 ScAlN 的不二法門，因為它會影響磊晶膜中缺陷的形成與結晶的品質，長晶時通常採用較高的 V/III 比值。

綜合言之：成長高品質 ScAlN 磊晶的關鍵為：需精準控制 N/(Sc + Al) 的比值，故在 MBE 系統中通常會採用較高的比值，亦即 N-rich 的狀況，因為表面動力學的研究顯示它可減少磊晶膜中缺陷的形成。此外介面應力的處理也很重要，因 ScAlN 具有較大的壓電係數會聚集很強的內應力導致其臨界厚度只約數個 nm，如 Sc ~ 15% 時就只有 5–8 nm，其對元件的影響就是造成閘極漏電流的增加，雖然如此但也不宜讓應力緩解，因為會使磊晶膜產生 crack 令元件在 RF 操作時雪崩式的失效，況且太薄的 ScAlN 也會使其 k 值降低，喪失其原本 high k 的優勢，導致 BV_{DS} 表現比 AlGaIn 更差，這些狀況都是 ScAlN 目前仍無法完全取代 AlGaIn 的最大瓶頸。因此欲製作高性能元件時對 ScAlN 能障層的要求需能達到下列目

標：(1) Sc 含量約介於 12–18%。(2) 厚度約 8–10 nm，但不能有緩解現象。(3) 介面粗糙度 RMS 需小於 0.5 nm。(4) 漏電流不可遜於 AlN 的水準。(5) 移動率需達 $1000 \text{ cm}^2/\text{V}\cdot\text{s}$ 以上。(6) 於 RF bias 時不得有 V_{th} 調控飄移或產生操作模式變換的極化切換 (polarization switching) 現象。

另外，成長 ScAlN 的方法雖多但各都有其罩門，如：MBE (PA-MBE) 雖擅長於 Sc 量的控制且有陡峭的介面，但因成長溫度較低易生成大量的點缺陷，而 MOCVD 雖宜於量產，但 Sc precursor 的選用是其最大的挑戰，更會衍生磊晶膜形成相分離與 Sc 分布不均勻的困擾，至於 Sputtering 則是難以達成元件品質的要求，因此可說目前並無「完美」的 ScAlN 磊晶成長法。但不論採用哪種成長技術都需謹記各種技術的瓶頸對元件特性的影響，如結晶相穩定與元件的 Sc 含量及極化特性有關，臨界厚度與元件的能障層厚度有關，界面粗糙度與元件的移動率和 f_T 有關，缺陷分布與元件的漏電流以及 BV_{DS} 相關，ScAlN 的介電損 (κ loss) 則與 g_m 與漏電流等有關。

2. 介面與接點技術

如前所述，介面問題起因於 Sc 併入 AlN 的晶體會導致結晶整體性破壞，劣化晶面的平滑度使 μ 降低限制了 f_T 的提升，也增大 RF 的雜訊，故 ScAlN/GaN 的表面及介面的改善是製作高性能 HEMT 的首要關鍵，而前面我們已詳述 ScAlN 長晶的重點，遵其方法應有助於介面平整問題的改善。

但介面的議題除平整度的要求外，HEMTs 元件的接點技術也很重要，因為 GaN 與 ScAlN 都是寬能隙的材料，而寬能隙常導致高 Schottky 能障，使歐姆接觸 (ohmic contact) 製作困難，若不處以重摻雜或介面修飾將難以達到 low specific contact resistivity (ρ_s) 的要求，也會使元件因工作於高 R_{on} 而造成高損耗。對元件的影響則是：未最佳化的 source/drain (S/D) 歐姆接觸會降低 HEMT 的輸出電流，限制高頻表現 (f_T , f_{MAX})，影響元件的效率與功率承受能力。尤其 ScAlN 能障層又具有很強的極化特性與甚高的 2DEG 密度，若 R_{on} 太大對其電流的表現將非常不利。至於在 Schottky 接觸造成的困擾則是影響閘極的控制，無法抑制 gate/drain 漏電流並劣化 BV 的表現。因此介面接點技術也是製作高性能 ScAlN HEMT 的重要關鍵。而相關的處理策略有：

(a) ScAlN 元件歐姆電阻的製作：一般 HEMTs 歐姆接點的傳統製法為鍍上金屬疊層 Ti/Al/Ni/Au 然後再以 RTA 熱處理，由於這方法需要較高的退火溫度 ($\geq 800^\circ\text{C}$) 故常造成緩衝層的劣化使可靠性變差，而且因金／半介面不平整也無法使 Rc 值大幅降低，限制了功率傳輸的效率。因此近期有文獻提出結合 Recess 與 Regrown n^+ GaN 的複合技術來製作歐姆接點，由於 Recess 縮短了金屬接點到 2DEG 的穿隧 (tunneling) 距離，又配合了在 S/D 的接點區以 MBE 或 MOCVD regrow 重摻雜的 n^+ GaN (or InGaN) 磊晶膜，因 regrown 區的歐姆接觸沒有經過退火處理 (non-alloyed) 因此介面會較平整，有效降低了 effective Schottky 能障令 Rc 大幅降低，這方法倍受肯定，有研究甚至以 Ge 取代 Si 作為重摻雜 n^+ 的雜質，藉由 Ge 摻雜在 regrown 區產生的壓縮應力 (相對於 Si 摻雜的張應力) 使 regrown GaN 磊晶膜更為平整晶膜缺陷減少，使 Rc 值更為降低，元件 RF 特性更佳。另為因應製作超薄 ScAlN barriers HEMT 時由於 ScAlN 薄化衍生的表面缺陷與粗糙性會影響接點電性，有文獻提出以 Tailored patterning 法來縮短 tunneling 的路徑以降低 Rc。總之，降低 Rc 的關鍵就在：儘量提高介面處的摻雜濃度並藉由 recess 及 regrown 法來降低接點／介

面能障的影響，以提高 thermionic emission (TE) 的效率。另有研究建議：製作陡峭的介面並減小 ScAlN 與 GaN 間應力，也能降低歐姆電阻⁽¹⁸⁾。

(b) 金屬閘 / ScAlN 元件 Schottky 接觸的製作：閘極漏電流是 HEMT 操作時最棘手的問題，因為它會使閘極的 AC 信號被分流 (shunt) 而降低 g_m ，也會消耗 AC 功率而損及電流增益，限制了 f_T 的提升。而且閘極功率被消耗也會造成 f_{MAX} 變差，甚至因功率損耗大使元件發熱觸發了增益壓縮 (gain compression) 提早發生，線性 (IP3) 因而變差，顯然閘極漏電流造成的麻煩還真多。那閘極漏電流的來源是什麼呢？有研究指出其主要來自：能障層中的缺陷密度，Dit，太高造成 trap-assisted tunneling (TAT)，以及閘極端點的電場太大造成 F-N 穿隧效應 (Fowler-Nordheim tunneling)，抑或因 barrier lowering 形成 poole-frenkel emission (P-F) 等因素。據觀察金屬閘 / ScAlN HEMT 的表現。發現 ScAlN 的能障 (Φ_B) 會因極化作用而受到調整 ($\Delta\Phi_{polarization}$)，可表為 $\Phi_{B,eff} = \Phi_B - \Delta\Phi_{polarization}$ ，顯然極化作用會使 Schottky 能障 (Φ_B) 降低同時窄化能障寬度，這效應增大了電子穿隧能障層的機會，令閘極漏電流大幅增加，而且強極化會使電場聚集在閘極端點 (gate edge)，使其也成為漏電流的熱點。再加上 ScAlN 特殊的材料因素，如 Sc 的併入使晶體致生各種缺陷，也形成了 tunneling 或 shunt 的路徑，造成其漏電流硬是較 AlGaIn 增大許多，這些亂象顯示了 ScAlN 能障層正是閘極漏電流的源頭，但必須強調：ScAlN 的漏電流太大並不是其材質不夠絕緣，而是極化作用造成電場分布失控再加上結晶缺陷雪上加霜地助長了穿隧漏電流所致，所以解決的關鍵不在更換閘極金屬材料，而是需要重塑能障層的能帶與電場的分佈。就此項金屬閘 / ScAlN HEMT 挑戰的解決文獻建議的 SOP 為：(1) 表面處理：ScAlN 因表面缺陷較 AlGaIn 更多，更活性也更容易氧化，故清潔處理程序不可少，須先以低損傷方式，如稀釋的 HF 或低功率的 ICP 清潔表面後立刻進行 *in-situ* 金屬蒸鍍。(2) 選擇適當的金屬：由於 ScAlN 介面缺陷特多，故需選擇不會與 Sc 或 Al 反應的金屬以免介面形成穿隧路徑，較佳的選擇有 Ni/Au 或 Pt/Au 等。(3) 最關鍵的 gate edge 電場重分布處理：由於 ScAlN 的強極化使其側向電場甚大於垂直方向使 gate edge 成為漏電流的熱點，解決的方法有：製作 T-gate，並插入高品質的閘極介質，如 Al_2O_3 (ALD) 或 *in-situ* 的 SiN，作為表面與 Sidewall 的保護披覆，以降低介面缺陷密度 Dit，接著製作閘極場板 (field plate) 以分散側向電場。(4) 控制閘極 recess 處理：以低功率的 ICP 處理 Schottky gate 以免造成電漿離子損傷衍生大量的缺陷，劣化 Schottky 的品質。據查經由上述步驟製成的元件目前較先進的成果已可將金屬閘 / ScAlN HEMT 的閘極漏電流壓低到 $10^{-6} - 10^{-7}$ A/mm，雖已較前改善惜仍未達 pico 安培等級⁽¹⁹⁾，而且漏電流還會隨偏壓增加而放大，顯然 ScAlN 強極化特性的陰影猶揮之不去。

(c) MIS 閘 / ScAlN 元件 Schottky 接觸的製作：基於金屬閘 / ScAlN 元件對閘極漏電流的改善有限，近年來高性能 ScAlN HEMT 的研究幾乎都改採 MIS 閘的結構，希望能更有效抑制閘極漏電流，提供更穩定的 RF 操作，如 f_T 與 f_{MAX} 等。至於閘極絕緣層的材料許多文獻都選用 Al_2O_3 /SiN 以有效阻斷電子的直接穿隧，之後輔以低能量的 ICP 蝕刻，然後進行 *In-situ* 的披覆保護以避免表面氧化，有些研究更加上 Recessed gate 的製程以調節能障層的厚度。這些製程確可將閘極漏電流抑至 $10^{-7} - 10^{-9}$ A/mm，但美中不足的是 f_T 僅達 50–55 GHz， f_{MAX} 也只有 70–75 GHz⁽²⁰⁾，顯然仍次於先進的金屬閘 / ScAlN HEMT 的表現，其在金屬微縮後典型的 RF 特性為 f_T 可達 80–110 GHz， f_{MAX} 甚至大於 100 GHz 的表現。分析原因可能與 ScAlN 介電材料的厚度以及介面品質不佳有關，影響 g_m 所致，

因為 MIS 閘會衍生寄生電容，如 C_{gs} ， C_{gd} ，由 (式 2) 可知其將劣化 f_T 的表現。然值得慶幸的是：閘極漏電流， I_g ，已可改善至 $10^{-7} - 10^{-9}$ A/mm，這優異的特性提供了 MIS 閘元件更大的電壓操作窗口，因為在高 V_{gs} 操作時低 I_g 可允許較大的 RF swing，因此提升了崩潰電壓的 margin，也可提供較穩定的 DC 工作點，又因為輸出功率是正比於 V_{swing} 的平方，所以可預期也將提供更高的輸出功率以及良好的線性度，讓 MIS 閘 / ScAlN HEMT 成為實用的毫米波功率放大器。惟目前 MIS 閘 ScAlN HEMT 的研究仍有限，難以論其功過，若欲確認其 RF 表現及應用的潛力仍需面對多項的挑戰。

- (d) ScAlN HEMT 製程的 ICP 技術：介面議題的解方猶須 ICP 技術的襄助。如前述，ScAlN HEMT 的閘極漏電流可藉 MIS 閘結構降低，但 MIS 結構會衍生較大的寄生電容 C_{gs} 反使 f_T 降低，這個 tradeoff 的結果正點出了 ScAlN HEMT 製程的一個關鍵，而其鎖鑰就在 ICP 技術。話說，在 ScAlN HEMT 製程中 ICP 不只做為尺寸的控制，也有調控側壁電場與管理陷阱源頭的任務。但須注意的是 ICP 蝕刻的物理行為會受 ScAlN 的影響而改變。下述說明其影響：(1) 在 AlGaIn HEMT 製程 ICP 多半作為圖形定義，但在 ScAlN 的製程它會直接影響漏電流、 f_{MAX} 與 current collapse 等特性，因為在 sidewall 蝕刻過程 ICP 角度的控制會改變閘極端電場分布的狀況，如 ICP 損傷層的厚度會影響 TAT 漏電流，ICP 造成的表面粗糙度會成為 FN tunneling 的熱點，甚至 ICP 殘留的氯化物也會成為表面漏電流的路徑，因此 ICP 在處理 ScAlN sidewall 時若控制不當將成為 HEMT 元件漏電流的禍首。(2) 藉由 ICP 對 sidewall 側壁角度的控制可決定閘極端的電場分布，如側壁形貌為垂直狀時會使場線集中造成漏電流增加，若改以微斜角如 $70 - 80^\circ$ 時場線會分散使漏電流減少，又若處以圓角 (rounded foot) 則電場的熱點可被移除。但須注意的是 ScAlN 極化特強對側壁角度頗為超敏感，故 ICP 控制處理時需仔細小心。(3) 以 ICP 製造平滑的側壁再加上適度的 recess 形貌，將可降低 Gate-to-Drain 的邊緣波動程度使 C_{gd} 減小，使 f_{MAX} 的表現提升，這個訣竅在 ScAlN 製程中特別重要，因為如前述 ScAlN 元件的 RF 特性常受累於偏大的 C_{gd} 使其難以提升。(4) ICP 中高能的離子會在 ScAlN 表面產生 N 空缺或 Sc 相關的缺陷，這些都將形成 TAT hopping center，故抑制之道就在控制 ICP 的能量以減輕電漿造成的損傷。(5) ICP 可使側壁平滑完整覆蓋 SiN/Al₂O₃ 保護層，改善其貼合度並阻斷表面漏電路徑。(6) 不同於 AlGaIn，ScAlN 會強烈影響 ICP 的蝕刻行為，因為 ScAlN 的 ICP 反應材料為 ScCl_x，其揮發性低容易在表面殘留非揮發性的化合物使表面粗糙度增加也降低了後續的蝕刻速率，故為加速蝕刻需得更提高離子能量但這又增加了元件受損的風險。(7) ScAlN 的 ICP 製程秘訣為：ICP 宜採低功率以減少物理傷害，並採用適當的氣體。總之，在 ScAlN/GaN HEMTs 的製程中以 ICP 蝕刻閘極不只是定義其幾何形狀，也擔負著需降低 sidewall 缺陷的生成與調節閘極端電場分布的重責大任，但因 ScAlN 會影響 ICP 蝕刻的動力行為 (etch kinetics) 使其表面損傷與粗糙程度較 AlGaIn。更為惡化，由於 sidewall 的損傷會增加元件的 C_{gd} ，閘極漏電流等特性，也增大了 RF loss，使 ScAlN/GaN HEMTs 的表現劣化，這些都是應用 ICP 時須特別注意的關鍵。

3. 高性能 ScAlN/GaN HEMTs 的製作設計

如前述，ScAlN 強烈的極化特性可將 2DEG 推到極高的境界使其大於 $2 - 3 \times 10^{13} \text{ cm}^{-2}$ ，提供元件甚大的輸出電流，這是其傲視其他材料的優勢，但這些電子並非全為「free」態，可貢獻於 f_T 的電子數其實是有限的，且強烈的 optical phonon 散射效應也會造成電子早發性

的速度飽和，甚且極化衍生的閘極電容等寄生效應，如 C_{gs} ， C_{gd} 等，也因偏高而不利於 f_T 的提升，顯然，ScAlN 強極化的優勢雖能載舟但也會覆舟，反而成為 RF 表現提升的障礙。另 ScAlN 材料又因化學特性極為不穩，合金化的接點製作容易劣化，Regrown 製作也不簡單，這些因素都令其歐姆接觸電阻偏高不下，也阻礙其可呈現高性能 HEMT 的期待。但據前述諸多研究的分析發現：接觸電阻的良窳對 ScAlN 元件 RF 表現的影響是遠大於極化的利益，所以在製作高 RF 性能的 ScAlN/GaN HEMTs 時若設計一道能「取其極化之利並去其弊」的製程就非常關鍵了，而且視 RF 應用不同，製程策略也應不同。

以低雜訊放大器 (LNA) 的製作而言，由於此 ScAlN HEMT 將應用於高速傳輸與需具備大頻寬的 5G FR2，(24.25 GHz-71 GHz)，毫米波系統以及 77–110 GHz 頻段等雷達與衛星前端的電路系統，故製作策略應為：將 ScAlN 元件推到以極化驅動的高密度與高速通道的境界將 f_T 大幅提升。但因電子的傳播行為會影響毫米波元件的 RF 表現，所以要使 ScAlN 元件的 f_T 表現優於 AlGaIn/GaN 的關鍵就在「極化工程」的設計，即藉由摻入 Sc 以增大 ScAlN 的極化作用使 2DEG 密度增加，再利用此超高極化電荷的優勢調變介面量子井使其變為更深、更陡的三角量子井，以提升通道電子的侷限能力。由 (式 1) 可看到 g_m 因 2DEG 的密度增加使分布變得更尖銳 (sharp peak)，而 g_m 極大值的出現意謂著 ScAlN 元件的 f_T 可被大幅提升。由於 f_T 表示元件可放大電流的能力，而此表現的 ScAlN HEMT 具備了高 g_m ，因而適合製作用於雷達與衛星前端電路系統的 LNA。

另就毫米波 / THz 功率放大器 (PA) 的製造，ScAlN HEMT 有個很重要的優勢，即在 f_T 達飽和後，其 f_{MAX} 仍可持續放大，由於 f_{MAX} 表示元件「可輸出功率增益」的最高頻率，故 ScAlN/GaN HEMT 這種甚高且穩定的 f_{MAX} 表現很適宜作為 5G 基地台，mmWave backhaul，衛星 uplink 與雷達發射端等尖端系統所需的 PA。又因 PA 的要求為高輸出阻抗，低 R_g 與大於 1 的 Mason's Unilateral power gain (Mason U) 使其可以提供穩定的淨功率增益，而這些要求都將影響 f_{MAX} 的表現。由前述的分析可知 ScAlN 元件 RF 特性的挑戰主要來自 C_{gd} 與 R_g 所主導的 Miller feedback 與損耗，而不是 g_m ，所以製作 ScAlN HEMT PA 的關鍵就在：讓有超高的極化特性的 ScAlN 產生高密度的 2DEG 以降低通道電阻提升 g_m 之後，再將結構優化，即利用電場重分布技術輔以阻抗工程技術，使 gate-channel 電容 (C_{gs})，gate 電阻 (R_g) 以及 feedback 電容 (C_{gd}) 等得以降低，並使場型集中於 gate-edge 使輸出阻抗 r_o 增大，並製作 T-gate 減小 R_g 以及利用 regrown contact 優化歐姆接觸與 access 區的電阻 (即 R_s 與 R_d)，由於上述製程可將造成損耗的寄生效應大致消除，則依 Mason U 係正比於 $(gain)^2 / (loss) (feedback)$ 的關係，U 值將可穩定維持在大於 1 的狀態。再由 (式 3) 可知 f_{MAX} 也將提升。前述 Nguyen 的元件就是依此策略使 f_T 飽和後， f_{MAX} 仍可持續提升，並將 f_{MAX} / f_T 推升到創紀錄的 321/ 173 GHz。惟這些方法都是針對製程考量，但已知 ScAlN 元件的 RF 特性會受其本質極化效應的限制而有天花板的極限其這缺憾是無法以製程消除的，況且 ScAlN 有些製程技術也尚未成熟，因此面對提升 ScAlN/GaN HEMTs 性能的挑戰猶須以更新的思維來考量，如適當控制極化效應但不能犧牲元件特性等方向。

GaN-基 HEMT 另有一關鍵技術就是場板 (field plate) 的設計，其目的為分散 gate edge 的電場，降低電荷捕獲，提高耐壓能力 V_{bd} ，與改善訊號的線性度。常見的設計為 gate field plate，它的優點在於可減少電流崩潰並提升 V_{bd} ，缺點則為使 C_{gd} 增加並降低 f_T 。另一種設計則為 source field plate，其優點為可抑制電荷捕獲效應衍生的 knee walkout 現象並改善功率附加效率 (PAE)，至於最佳的方案則為。Dual field plate，它擁有較高的 PAE 與 V_{bd} 且可接受的 f_T 。當然上述各種製作方案也得面對處理 ScAlN 超強的極化特性的挑戰。

七、結論

近年來有關 ScAlN 的研究都認為其材料特性較 AlGaIn 有更多的優點，有取而代之的潛力。然而在評比兩者的 RF 表現時卻發現 ScAlN HEMT 在一些重要特性上，如 f_T 與 f_{MAX} ，仍遠遜於其他的 GaN-基元件，其原因可能是前者尚未建立如後者般成熟的技術，若改善製程應可迎頭趕上，故本文探討了現有 GaN-基元件先進 RF 表現的製作細節，並盤點製作高性能 ScAlN/GaN HEMTs 的幾點關鍵技術，如：(1) ScAlN 磊晶膜的成長。(2) 接點技術與 gate 漏電流的抑制。(3) ScAlN/GaN 元件的製作設計策略，期能助於提升 ScAlN 元件特性。惟觀察近期 ScAlN 的研究，其 HEMT 特性雖已獲得部分改善，但仍有缺憾，如元件的 mobility 偏低，gate 漏電流太大限制了 f_T 的提升，介質介面有加速型 TDDDB 的劣化。而且元件的 V_{th} 有飄移現象，這些困擾都成為 ScAlN/GaN 元件發展的隱憂。研究發現這些困擾的源頭就在 ScAlN 引以為傲的本質極化特性，它使 ScAlN/GaN HEMT 的 RF 的表現罩上一個天花板的限制，這個本質的問題令 ScAlN 元件的製作面臨了一個非製程可解決的挑戰。

考量到水能載舟也能覆舟，因此製作高性能 ScAlN HEMT 時不能只單純的更換能障層材料而應以的新思維，如善加利用極化特性，發展電場重分布技術與阻抗工程技術，將之與元件結構、寄生效應等參數耦合考量，使 ScAlN/GaN HEMT 達到以極化驅動的高密度高速通道的狀態來提升 f_T ，這個策略將適於 LNA 的製作。惟若耦合效果是使元件達到高輸出阻抗、低通道電阻的境界，且可令 Mason U 穩定維持大於 1 使得以持續產生功率增益的狀態，則將讓 f_T 由速度主導，而 f_{MAX} 則由阻抗與寄生效應主導，呈現最理想的毫米波放大器特性，這個策略則是製作 ScAlN/GaN HEMT PA 的關鍵。總之，ScAlN 具備相當的製作彈性，這是它可應用於下世代尖端高頻高性能電子系統的最大優勢。

當然 RF 應用所需的特性除 f_T 與 f_{MAX} 外還有：PAE (power added efficiency)，Pout、Gain、Linearity 與 Breakdown/reliability 等重點，但其改善策略則和 f_T 、 f_{MAX} 不同。尤其目前文獻大都認為 gate leakage、dynamic R_{ds} 與 thermal reliability 三者是 ScAlN 元件開發的瓶頸，ScAlN/GaN HEMT 欲成為毫米波 PA 的核心技術需同時滿足 f_T 大於 200 GHz， f_{MAX} 大於 300 GHz 與 PAE 大於 70% 的要求，然達成這個目標的障礙就在極化特性的處理，顯然 ScAlN 元件的研發不能只單純的視其為較強的 AlGaIn，而須嚴肅看待極化特性對其元件表現的影響。

參考資料

1. J. Wuerf *et al.*, *Microelectronics Reliability*, **51**, 1710 (2011).
2. 洪茂峰, 科儀新知, 245期, 12月, 79 (2025).
3. Peng Cui and Yuping Zeng, *Scientific Reports*, **12**, 16683, (2022).
4. Shiming Li *et al.*, *Nanomaterials*, **15** (24), 1882 (2025).
5. Wang *et al.*, *Chinese Physics B*, **29** (2), 027301(2020).
6. Liuyun Yang, *ACS Appl. Electron. Mater.*, **5** (9), 4786 (2023).
7. J. Cheng *et al.*, *IEEE Electron Device Letters*, **42** (8), 1132 (2021).
8. S.Krause *et al.*, *IEEE Electron Device Letters*, **44** (1), 17 (2023).
9. Lei Li *et al.*, *IEEE Electron Device Letters*, **41** (5), 689 (2020).
10. K. Nomoto *et al.*, *Applied Physics Express*, **18**, 016506 (2025).
11. Thai-Son Nguyen, K. Nomoto, Wenwen Zhao, Chandrashekhar Savant, 2024 IEEE International Electron Devices Meeting (IEDM), December 7-11 (2024).
12. A. Hickman *et al.*, *IEEE Electron Device Letters*, **40** (8), 1293 (2019).

13. Dambrine *et al.*, *IEEE Transactions on Microwave*, **36** (7), 1151 (1988).
14. Okuda *et al.*, *Appl. Phys. Lett.*, **126**, 052105 (2025).
15. Casamento *et al.*, *Appl. Phys. Lett.*, **20**, 152901 (2022).
16. Žukauskaitė *et al.*, *Acta Materialia*, **94**, 101 (2015).
17. A.Tomita *et al.*, *Scientific Reports*, **13**, 3308 (2023).
18. Isabel Streicher *et al.*, *Advanced Functional Materials*, **34** (39), (2024).
19. Jiangnan Liu *et al.*, *Appl. Phys. Lett.*, **126** (25), 253505 (2025).
20. Su, Chin-Ya *et al.*, 2024 International VLSI Symposium on Technology, Systems and Applications (VLSI TSA), April 22-25, (2024).
21. Chumbes *et al.*, *IEEE EDL* 20, **12**, 630 (1999).

作者簡介

洪茂峰先生為成大電機博士，現任職成大電機系兼任教授。

Mau-Phon Houng received his Ph.D. in EE from NCKU, Tainan, Taiwan. He is currently a Adjunct professor of Dept. of EE, NCKU.