

淡江大學電機工程學系專題

指導教授：江正雄博士

**Swin TransKalman 具不確定性量化之卡
爾曼濾波解碼器於腹部多器官分割**

**A Kalman Style Decoder with
Uncertainty Quantification for
Abdominal Multi Organ Segmentation**

學生：詹盛惟

潘信瑞

沈群博

中華民國 115 年 10 月

目錄

摘要	1
Abstract	2
第一章 緒論	3
1.1 研究背景	3
1.2 多尺度融合的研究問題	3
1.3 研究目的與貢獻	4
1.4 研究範圍與論文架構	5
第二章 相關文獻與理論基礎	6
2.1 三維卷積與 U 形分割架構	6
2.2 自注意力與階層式轉換器	6
2.3 轉換器於醫學影像分割的方法	7
2.4 卡爾曼濾波的預測與更新	8
2.5 可學習濾波與結構化神經模組	9
2.6 不確定度估計與錯誤辨識	10
2.7 深監督與資源受限的訓練	11
2.8 評估指標與本文定位	11
第三章 研究方法	13
3.1 整體架構與問題定義	13
3.2 具正值狀態傳播的解碼器	14
3.3 編碼器的卡爾曼式閘門	16
3.4 注意力的溫度與狀態偏置	17
3.5 不確定度讀出與訓練目標	19
3.6 訓練與推論程序	21
3.7 方法特性與可檢驗假說	22
第四章 實驗結果與討論	23
4.1 資料集與交叉驗證設計	23
4.2 前處理與實驗設定	24

4.3 評估指標與統計方法	25
4.4 五折整體分割結果	27
4.5 逐折與逐案例分析	28
4.6 逐器官表現與失敗模式	31
4.7 深監督的補充消融	33
4.8 不確定度的錯誤辨識能力	34
4.9 收斂行為與視覺化觀察	36
4.10 綜合討論與研究限制	37
第五章 結論與未來工作	39
5.1 研究結論	39
5.2 研究貢獻與適用範圍	40
5.3 未來工作	40
參考文獻	41

圖目錄

圖 1 正值狀態與觀測雜訊對融合增益的影響	9
圖 2 Swin-TK 整體架構與雙輸出路徑	13
圖 3 卡爾曼式解碼器的預測、觀測與更新流程	16
圖 4 編碼器中閘門、注意力調節與對數狀態的關係	18
圖 5 查詢溫度與鍵位置偏置的作用示意	19
圖 6 五折交叉驗證的訓練與驗證分配	23
圖 7 五折主實驗的重疊與表面重疊係數比較	28
圖 8 SwinUNETR 與 Swin-TK 的逐折重疊係數	29
圖 9 表面評估流程中逐案例的重疊與表面重疊係數配對	30
圖 10 各器官的重疊與表面重疊係數平均差值	32
圖 11 第 1 折的深監督補充消融結果	34
圖 12 不確定度與影像梯度的逐折錯誤辨識能力	35
圖 13 第 1 折訓練期間的驗證重疊係數曲線	36
圖 14 開發階段示例的電腦斷層影像、分割、標註與不確定度	37
圖 15 開發階段示例的錯誤位置與高不確定度區域	37

表目錄

表 1 研究問題與對應證據.....	5
表 2 相關分割方法的設計方向.....	8
表 3 不確定度估計途徑與本文評估範圍.....	11
表 4 五層解碼器的張量配置.....	14
表 5 訓練與推論流程.....	21
表 6 資料與交叉驗證範圍.....	24
表 7 各折訓練資料的電腦斷層影像正規化參數.....	24
表 8 模型、訓練及推論的主要設定.....	25
表 9 正規化表面重疊係數的逐器官距離容忍值.....	26
表 10 五折完整方法與基準模型的比較.....	28
表 11 主要重疊係數的逐折比較.....	29
表 12 主要重疊係數的案例配對分析.....	30
表 13 表面評估流程的逐器官重疊係數.....	31
表 14 表面評估流程的逐器官表面重疊係數.....	32
表 15 第 1 折深監督補充比較.....	34
表 16 不確定度與影像梯度的逐折錯誤辨識能力.....	35
表 17 主要研究主張與目前證據.....	38

摘要

腹部多器官分割需同時整合粗尺度語意與細尺度邊界，但一般跳接融合未顯式傳播來源可信程度，亦難以由單一分割結果辨識容易出錯的位置。本研究提出 Swin-TransKalman (Swin-TK)，以 SwinUNETR 為基礎，將解碼過程表述為沿尺度的預測與更新。模型以可學習正值狀態及觀測雜訊形成逐元素增益，融合解碼特徵與編碼跳接；更新狀態持續向高解析度傳遞，其摘要參與特徵精煉，並由最終狀態讀出逐體素不確定度。編碼器另整合卡爾曼式閘門、注意力溫度與對數狀態偏置，使資訊選擇與特徵更新共同參與分割學習。研究使用 FLARE22Train 的 50 例標註電腦斷層影像，分割 13 類腹部器官及背景，採五折內部交叉驗證，每折以 40 例訓練、10 例驗證。兩模型皆使用 1,000 個訓練週期，並以完整體積滑動視窗推論取得分割結果；完整 Swin-TK 另使用深監督與不確定度損失。完整 Swin-TK 的平均重疊相似係數 (DSC) 由基準模型的 0.8649 提升至 0.8755，正規化表面重疊係數 (NSD) 由 0.8917 提升至 0.9063，分別增加 1.06 與 1.46 個百分點，參數量由約 6,220 萬增至 6,360 萬。五折重疊係數皆呈正向差值，主評估中 39/50 例改善。第 1 折的補充實驗顯示，兩模型皆未使用深監督時，Swin-TK 仍由 0.8706 提高至 0.8769；加入深監督後達 0.8808。不確定度 σ 的五折接收者操作特徵曲線下面積 (AUROC) 為 0.776 ± 0.039 ，高於影像梯度的 0.609 ± 0.041 。結果支持所提出結構化融合在目前資料與訓練條件下改善整體分割，並提供具錯誤排序能力的輔助輸出。單折消融支持提升不完全依賴深監督，但尚不足以分配各模組的獨立貢獻；不確定度結果亦不等同於機率校準。後續可透過完整模組消融、多種子訓練及外部資料驗證，進一步檢查方法的穩定性與適用範圍。

關鍵詞：腹部多器官分割、SwinUNETR、卡爾曼式融合、不確定度、深監督

Abstract

Abdominal multi-organ segmentation requires the integration of coarse semantic context and fine spatial detail. This thesis proposes Swin-TransKalman (Swin-TK), a SwinUNETR-based model with learnable Kalman-style feature fusion. Its decoder propagates positive states across scales and uses elementwise gains to combine predicted features with skip observations. State summaries also enter feature refinement, while a learned readout produces voxelwise uncertainty. The encoder incorporates gated feature updates, query-dependent attention temperatures, and propagated log-state biases. Five-fold internal cross-validation was conducted on 50 labeled CT volumes from FLARE22Train, covering 13 abdominal organs and background. Each fold used 40 volumes for training and 10 for validation. Both models were trained for 1,000 epochs and evaluated with full-volume sliding-window inference; the complete Swin-TK additionally used deep supervision and an uncertainty objective. The complete method increased mean Dice from 0.8649 to 0.8755 and normalized surface Dice from 0.8917 to 0.9063, while the parameter count increased from approximately 62.2 M to 63.6 M. Dice improved in all five folds and in 39 of 50 cases in the primary evaluation. In a supplementary fold-1 experiment without deep supervision for either model, Swin-TK achieved 0.8769 compared with 0.8706 for the baseline. Enabling deep supervision increased its Dice to 0.8808. The learned uncertainty achieved an error-detection AUROC of 0.776 ± 0.039 across folds, compared with 0.609 ± 0.041 for image gradients. These results support improved segmentation and useful error ranking under the evaluated conditions. The single-fold ablation shows that the gain does not depend entirely on deep supervision, but does not isolate every module. The uncertainty scores are evaluated as ranking signals rather than calibrated error probabilities. Further module ablations, repeated training, and external evaluation are needed to establish broader robustness.

Keywords: abdominal multi-organ segmentation; SwinUNETR; Kalman-style fusion; uncertainty; deep supervision.

第一章 緒論

1.1 研究背景

腹部電腦斷層影像的多器官分割，是將三維影像中的肝臟、腎臟、脾臟、胰臟與血管等結構轉換為逐體素標記的過程。這項工作除了需要辨識器官的位置，也必須在連續切片間維持形狀與邊界的一致性。對自動分割系統而言，大型器官的內部區域通常具有較充足的影像證據，細小器官、器官接觸面與對比不明顯的邊界則較容易產生混淆。因此，模型的整體重疊分數與局部錯誤分布，分別反映系統表現的不同面向；只觀察平均分數，往往不足以說明錯誤發生的位置與性質。U 形編碼器與解碼器架構已成為醫學影像分割的重要設計。U-Net 利用下採樣路徑取得較大範圍的語意資訊，並透過跳接將較高解析度的特徵傳入上採樣路徑，以恢復空間細節[1]。3D U-Net 與 V-Net 進一步將體積資訊納入特徵運算，使模型能同時利用切片內與切片間的關係[2,3]。這些方法說明，多尺度特徵的取得與融合，是三維分割架構的重要組成。當解析度下降時，特徵具有較廣的上下文，但部分細節也會被壓縮；當解析度提高時，局部邊界較清楚，卻可能缺少判定器官身分所需的整體結構。

Transformer 將資訊交換建立在特徵之間的相似度上[5]。在視覺任務中，階層式設計與區域注意力降低了高解析度輸入的運算負擔，Swin Transformer 以視窗及位移視窗建立局部與跨視窗的關聯[7]。SwinUNETR 結合三維 Swin 編碼器與卷積解碼器，提供多尺度的影像表徵與分割輸出[9]。本研究採其作為基礎，研究重點在於編碼與解碼過程如何使用來自不同位置及不同尺度的資訊，而非重新建構一套完全獨立的特徵擷取骨幹。分割模型的輸出通常可轉換為類別機率，但機率最大值與實際預測可靠度之間並不必然一致。模型可能在容易辨識的區域給出明確判斷，也可能在錯誤區域維持高信心。相關研究因此將不確定度估計視為預測系統的重要補充，探討如何從模型參數、輸出分布或額外估計量中取得可信度資訊[16–19]。對本研究而言，不確定度的目的不是取代器官分割，而是在保留分割功能的前提下，提供與錯誤位置相關的連續訊號，協助分析模型在哪些區域較容易失準。

1.2 多尺度融合的研究問題

標準跳接融合通常將上採樣特徵與編碼器特徵串接，再透過卷積及非線性轉換取得新的表示。這種設計具有良好的可學習性，卷積網路也能依輸入內容產生不同反應，但融合過程沒有顯式保存「目前估計狀態」與「新觀測可信程度」兩

種表示。因此，當粗尺度的器官語意與細尺度的局部邊界資訊不一致時，模型主要透過整體權重學習處理這種差異，較難直接描述每一層如何分配兩種來源的影響。本研究將解碼過程視為沿尺度逐步更新特徵狀態的程序。較深層的特徵提供目前的粗略估計，跳接特徵提供當前尺度的觀測，兩者經過通道與解析度對齊後，以位置相關的融合增益完成更新。這個觀點與卡爾曼濾波的預測與更新形式相容[14]，但影像網路中的狀態不是具有已知物理單位的動態系統變數，特徵轉移及雜訊投影亦由訓練決定。因此，本文使用卡爾曼式更新作為結構化設計原則，並透過分割實驗檢驗其效果。

另一個問題是編碼器內部的資訊交換。注意力權重由查詢與鍵的相似度產生，同一個查詢會根據多個來源建立加權表示。注意力溫度能改變分布的集中程度，而逐來源的加性偏置可以直接改變不同鍵之間的相對權重。本研究在編碼器中結合這兩種調節方式，並加入沿深度傳遞的純量狀態，使每一個區塊除了更新特徵，也保留可供下一個區塊使用的調節資訊。此設計與解碼器的跨尺度傳播具有共同的更新概念，但兩者的狀態維度、傳播軸與使用方式均有所不同。若新增的狀態只用於產生視覺化圖片，其與分割過程的關係仍然薄弱。本研究因此讓解碼器的不確定度摘要參與後續特徵細化，並在最後由傳播狀態讀出逐體素估計量。這使不確定度分支與特徵融合具有結構上的連結。然而，結構連結本身不足以保證輸出具有錯誤辨識能力，仍須將不確定度分數與實際分割錯誤配對評估。本文因而同時處理分割品質與錯誤辨識兩個問題，而不以其中一項指標替代另一項。

1.3 研究目的與貢獻

本研究提出 Swin-TransKalman，以下簡稱 Swin-TK，研究目標是在既有 U 形 Transformer 中建立可學習的卡爾曼式融合與不確定度傳播，並於腹部多器官分割中評估其效益。架構設計包含解碼器的跨尺度預測與更新、編碼器的殘差門控及注意力調節，以及由傳播狀態讀出的不確定度圖。各模組均保留可微分運算，使分割與不確定度相關參數能在共同訓練程序中最佳化。第一項貢獻是將解碼器的跳接融合表述為狀態與觀測之間的逐元素更新。模型以正值狀態和可學習的觀測雜訊形成增益，依位置與通道調節粗尺度特徵及跳接資訊。更新後的正值狀態持續傳至下一個尺度，其通道摘要亦被送入特徵細化層。此設計使各層的融合決策具有一致的計算形式，並將不確定度表示納入解碼過程。

第二項貢獻是整合編碼器殘差門控、注意力溫度及沿深度傳播的把握度偏置。殘差門控控制舊表示與區塊輸出之間的融合，溫度調節影響查詢對鄰域的權

重集中程度，逐鍵偏置則提供調整資訊來源相對重要性的途徑。三者分別作用於特徵更新、注意力分布形狀與來源選擇，構成完整 Swin-TK 的編碼器設計。第三項貢獻是以交叉驗證及補充消融評估整體方法。五折主實驗的平均 Dice 由 0.8649 提升至 0.8755，NSD 由 0.8917 提升至 0.9063。於第 1 折的補充比較中，兩模型皆未使用深監督時，Swin-TK 的 Dice 為 0.8769，高於基準模型的 0.8706；加入深監督後則達到 0.8808。不確定度評估中， σ 的五折 AUROC 為 0.776 ± 0.039 ，高於影像梯度的 0.609 ± 0.041 。上述結果分別支持整體分割改善、未使用深監督時的正向表現，以及輸出不確定度的錯誤辨識能力。

表 1 研究問題與對應證據

研究問題	方法設計	評估方式
融合能否改善分割	跨尺度 P 與卡爾曼式增益	五折重疊、表面重疊係數與案例配對
改善是否依賴深監督	關閉 Swin-TK 輔助分割損失	第 1 折三組設定比較
輸出能否辨識錯誤	由最終 P 讀出 σ	不確定度與影像梯度的逐折錯誤辨識能力

1.4 研究範圍與論文架構

本文使用 FLARE22Train 的 50 例標註 CT，分割目標為 13 類腹部器官，另以背景構成第 14 個輸出類別。資料以固定隨機種子切分為五折，每折使用 40 例訓練及 10 例驗證。主實驗限於標註資料的監督式訓練，不納入未標註資料、測試時增強、模型集成或額外後處理。此範圍有助於聚焦所提出方法在既定資源與訓練程序下的效果，並避免將不同資料使用方式的結果混合解讀。本文的比較分成兩個層次。五折主實驗比較基準模型與完整 Swin-TK 的整體表現，完整方法包含深監督與不確定度相關訓練。補充實驗則於第 1 折關閉 Swin-TK 的深監督，以觀察該設定下是否仍具有分割改善。前者評估完整方法，後者提供訓練因素的局部分析；單折差值不被解讀為各模組在所有資料切分上的固定貢獻比例。對不確定度的研究則聚焦分割錯誤排序，尚不將輸出數值視為經校準的錯誤機率。

全文共分五章。第一章說明研究問題、目的與範圍。第二章回顧三維分割、Transformer、卡爾曼式更新及不確定度估計，建立方法設計所需的觀念。第三章完整說明 Swin-TK 的架構、狀態傳播、損失函數、訓練及推論流程。第四章呈現資料與實驗設定、五折結果、逐器官分析、深監督消融及不確定度評估，並討論結果所支持的結論。第五章總結研究成果與適用範圍，提出後續可延伸的驗證與應用方向。

第二章 相關文獻與理論基礎

2.1 三維卷積與 U 形分割架構

語意分割需要在每個空間位置產生類別判斷，因此模型同時面對語意辨識與位置恢復兩項要求。U-Net 以收縮路徑擷取上下文，以擴張路徑回復解析度，並將相同尺度的特徵透過跳接連結[1]。從資訊處理的角度觀察，較深的特徵有助於辨識結構所屬類別，較淺的特徵則提供位置及紋理。解碼器的主要工作因此不是單純放大特徵圖，而是在較高解析度上重新整合不同層級的表示。3D U-Net 將二維運算延伸至三維，利用相鄰切片共同描述體積結構[2]；V-Net 則以三維卷積與體積分割損失處理類別區域的重疊問題[3]。對腹部 CT 而言，三維運算能提供單張切片以外的連續性，但其記憶體需求也會隨空間維度增加。實際訓練通常需在輸入範圍、特徵通道數與批次大小間取捨。若輸入區塊過小，模型可能缺少器官相對位置的資訊；若區塊過大，則可能限制可使用的網路深度與訓練批次。

nnU-Net 強調前處理、資料尺度、架構與訓練程序的整體配置[4]。這個觀點對新架構的評估具有直接影響：分割結果的變化可能同時來自特徵設計及訓練條件，因此比較時必須明確列出資料切分、取樣方式、最佳化器與損失。本文以固定的資料與訓練預算建立基礎比較，再以額外的無深監督設定分析部分差異；並不將採用 nnU-Net 風格的訓練設定，等同於完整重現 nnU-Net 自動配置流程。跳接中的串接與卷積是一種通用且有表達能力的融合方式，不能將其描述為完全不具位置適應性。卷積權重雖在空間上共享，輸入相關的非線性反應仍可產生不同位置的輸出。本文關注的是另一項設計差異：是否將來源可信程度明確表示為可傳遞的狀態，並以此建立每一層的融合增益。此問題聚焦融合的結構與可分析性，而不是否定傳統跳接的表示能力。

2.2 自注意力與階層式轉換器

自注意力將輸入特徵投影為查詢、鍵及值，先以查詢與鍵的內積計算相關性，再以正規化權重組合值向量[5]。若一個視窗包含 N 個 token，每個注意力頭的通道維度為 d ，則相似度矩陣具有 $N \times N$ 個元素。縮放因子可控制內積隨維度增加而擴大的幅度，softmax 則使每個查詢對所有鍵的權重和為一。其基本形式如式 (1) 所示。注意力的資料相依性來自查詢與鍵的內容，因此同一組模型參數可在不同影像上建立不同的關聯。另一方面，權重較大僅代表該來源在目前運算中受到較多使用，不能直接等同於該來源的分類正確率。注意力權重、特徵幅度與預測誤差描述不同的量，若要將其中某一訊號作為可靠度指標，仍須透過明確

的評估定義建立關聯。本文據此分開處理注意力調節的運算功能與輸出不確定度的錯誤辨識能力。

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V \quad (1)$$

Vision Transformer 將影像切成區塊並建立 token 表示，說明 Transformer 可應用於視覺辨識[6]。然而，對三維體積直接使用全域注意力，會使 token 數量及相似度矩陣快速增加。Swin Transformer 以階層式解析度與區域視窗控制運算規模，透過相鄰區塊的視窗位移建立跨視窗資訊交換[7]。若每個視窗含 M 個 token，視窗內的注意力運算約隨視窗數與 M 的平方成長，而非對完整體積建立單一全域矩陣。階層式編碼器能同時產生多個解析度的特徵，適合與解碼器的逐層恢復結構搭配。較淺 stage 保留細緻空間分布，較深 stage 以較少 token 表達較大的區域。這種表示適合三維分割，但不同解析度的 token 並非直接具有相同的語意尺度。若額外引入一張調節圖，必須處理它與特徵在切窗、位移及下採樣後的對齊，否則原本用來控制資訊來源的訊號可能落在錯誤位置。第三章將說明本研究的狀態傳播如何與特徵變換保持一致。

2.3 轉換器於醫學影像分割的方法

TransUNet 將 Transformer 與 U 形解碼器結合，以卷積與注意力共同處理醫學影像表示[10]。UNETR 則以 Transformer 編碼器取得三維 token 特徵，並將不同深度的表示連至解碼器[8]。這些研究反映了共同的方向：注意力負責建立較廣範圍的關聯，解碼器則回復空間資訊並產生密集預測。對本研究而言，相關工作的啟示是保留成熟骨幹，將設計焦點放在資訊如何進入解碼器與參與融合。SwinUNETR 使用階層式 Swin 編碼器，其多尺度輸出可與不同解析度的解碼特徵對應[9]。nnFormer 探討三維分割中的交錯式注意力與不同範圍的關聯建模[11]。相較之下，Attention U-Net 以注意力開調節跳接特徵，使解碼器更聚焦於任務相關區域[12]。這些方法的機制並不相同，但都指出特徵的傳遞與選擇，會影響模型對局部細節及整體結構的利用方式。

近年的 U-Mamba 將狀態空間模型與卷積整合，探討長距離關聯與局部特徵之間的搭配[13]。狀態空間模型與本文的卡爾曼式解碼器均使用狀態更新的語言，但兩者的目的與實作不同：前者主要作為序列或特徵建模元件，本文則在既有多尺度架構中，以正值狀態調節融合增益。本文不將使用狀態更新形式視為單一研究路線，而是根據各方法實際處理的問題與運算位置進行比較。表中整理的是設計方向，不是相同條件下的性能排名。不同文獻使用的資料集、標註比例、輸入範圍與預訓練條件可能不同，直接比較其 Dice 容易混入架構以外的因素。本

文的定量基準因此以相同內部資料切分上的 SwinUNETR 為主；其他方法用於界定設計脈絡與可能的比較方向，不將其公開分數併入本文的五折結果表。

表 2 相關分割方法的設計方向

方法	主要設計	與本文的關係
U-Net/3D U-Net [1,2]	多尺度編碼、跳接與解碼	提供 U 形融合基礎
nnU-Net [4]	資料相依的流程配置	強調架構以外的實驗條件
TransUNet [10]	卷積與轉換器結合	結合局部及上下文表示
UNETR [8]	三維轉換器編碼器	將特徵單元表示接至解碼器
SwinUNETR [9]	階層式視窗注意力	本研究的基礎骨幹
nnFormer [11]	三維交錯式注意力	研究注意力與多尺度表示
Attention U-Net [12]	跳接特徵的注意力閘	與來源選擇的問題相關
U-Mamba [13]	卷積與狀態空間模組	另一種狀態建模方向

註：本表為設計脈絡整理，未進行跨文獻分數排名。

2.4 卡爾曼濾波的預測與更新

卡爾曼濾波以狀態方程與觀測方程描述系統演化，將上一時刻的估計向前推進，再利用新觀測修正預測[14]。在常見的線性形式中，狀態轉移矩陣描述狀態如何變化，觀測矩陣描述狀態與量測的關係，過程及觀測雜訊則描述兩種資訊來源的不確定程度。其基本價值在於把「目前估計」與「新證據」放在同一更新框架中，而不是直接捨棄其中一方。在預測步驟中，估計狀態經由轉移矩陣更新，共變異數亦經對應的矩陣運算與過程雜訊累加。觀測到來後，卡爾曼增益決定創新量對狀態修正的影響。若觀測被認為較可靠，更新可以較大程度依賴觀測；若觀測雜訊較大，則較多保留原先預測。此處的可靠性由模型假設與雜訊描述共同決定，並非由增益公式單獨產生。

$$\mathbf{x}_t = \mathbf{F}\mathbf{x}_{t-1} + \mathbf{w}_t, \quad \mathbf{z}_t = \mathbf{H}\mathbf{x}_t + \mathbf{v}_t \quad (2)$$

$$\begin{aligned} \hat{\mathbf{x}}_t^- &= \mathbf{F}\hat{\mathbf{x}}_{t-1} \\ \mathbf{P}_t^- &= \mathbf{F}\mathbf{P}_{t-1}\mathbf{F}^T + \mathbf{Q} \end{aligned} \quad (3)$$

$$\mathbf{K}_t = \mathbf{P}_t^- \mathbf{H}^T (\mathbf{H}\mathbf{P}_t^- \mathbf{H}^T + \mathbf{R})^{-1} \quad (4)$$

$$\begin{aligned} \hat{\mathbf{x}}_t &= \hat{\mathbf{x}}_t^- + \mathbf{K}_t (\mathbf{z}_t - \mathbf{H}\hat{\mathbf{x}}_t^-) \\ \mathbf{P}_t &= (\mathbf{I} - \mathbf{K}_t \mathbf{H}) \mathbf{P}_t^- \end{aligned} \quad (5)$$

若將觀測模型簡化為單位映射，並只保留逐元素的正值狀態，增益可寫為 P 除以 P 與 R 的和。對正值 P 與 R 而言， K 介於零與一，更新可改寫為兩個來源的加權組合。這個代數性質提供了可控制的融合範圍：當 K 接近零，結果接近原先預測；當 K 接近一，結果接近觀測。圖 1 以不同 P 值展示 R 變化時的增益曲線，數值僅用於說明公式，並非模型實驗結果。將卡爾曼形式用於深度特徵時，必須注意理論假設的差異。卷積與非線性層不同於已知線性狀態轉移，通道之間也可能具有未被對角表示捕捉的關聯。因此，本文將 P 視為受到正值約束、沿尺度傳播的可學習狀態，使用其形成融合增益；不主張該張量在所有條件下都是具有完整機率意義的真實共變異數。這種界定使方法的可實作性與數學語意維持一致。

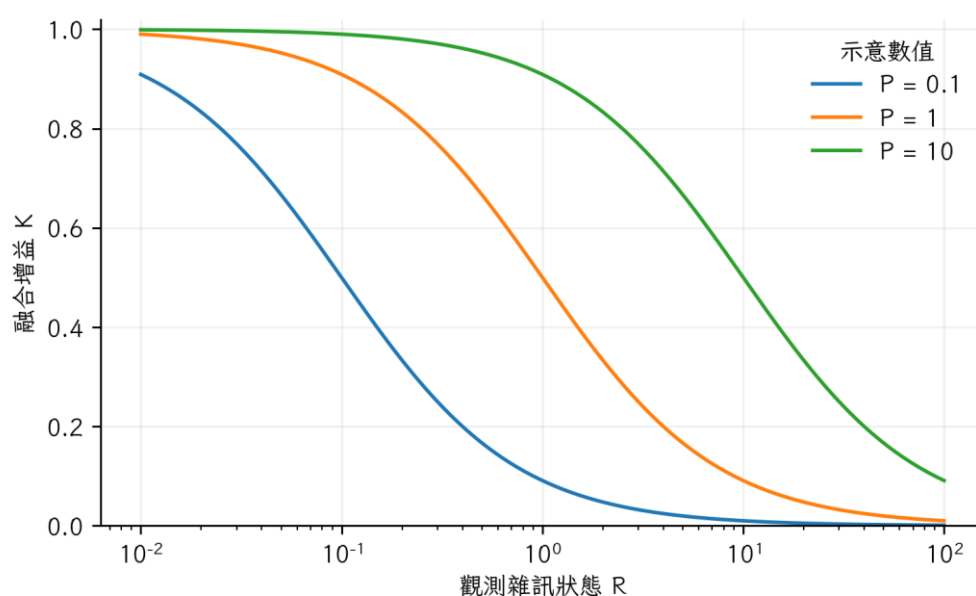


圖 1 正值狀態與觀測雜訊對融合增益的影響

2.5 可學習濾波與結構化神經模組

當系統動態或雜訊模型無法完全取得時，可學習模組可用於補足傳統估計器所需的部分映射。KalmanNet 將神經網路與卡爾曼濾波結合，研究部分已知動態下的狀態估計[15]。這類方法顯示，保留更新結構並讓部分量由資料學習，是可行的設計方向；但其有效性依賴特定任務、輸入與訓練方式，不能直接移植其估計保證至影像分割。本文與此方向的連結在於使用結構化更新形式，但狀態不是物體位置或速度，而是網路特徵與其輔助表示。解碼器的尺度軸也不同於物理時間軸。這些差異會影響 P 、 Q 與 R 的解讀：它們在本文中主要控制融合與表示傳播，而不是由已知感測器精度直接設定。模型參數透過分割及不確定度相關損失學習，最終效果由影像分割結果與錯誤辨識指標檢驗。結構化模組的價值不只在

於新增參數，也在於限定資訊可以如何互動。例如，正值轉換限制增益的輸入範圍，差值更新使兩種來源的修正關係明確，跨尺度狀態則提供比單層門控更長的傳遞路徑。然而，這些設計理由屬於方法動機，不能代替消融實驗。本文因而將設計上的預期與實驗中已觀察到的結果分開敘述，避免由形式相似性直接推論性能優勢。

2.6 不確定度估計與錯誤辨識

不確定度估計可涉及資料本身的模糊性，也可涉及模型參數或知識不足造成的變異。Kendall 與 Gal 討論影像任務中的偶然不確定度及認知不確定度，並提出相應的學習方式[16]。兩者的差異提醒我們，不確定度不是單一且具有固定解釋的數值。對一個實際模型而言，必須先說明估計量如何產生，再界定它與預測誤差、分布變化或機率校準之間的關係。Monte Carlo dropout 透過推論時的隨機遮罩產生多次輸出，以近似描述模型的不確定性[17]；深度集成則使用多個模型的輸出差異建立預測分布[18]。這些方法利用重複預測取得變異資訊，適合研究模型間或參數近似所造成的差異。本文選擇在一次網路前向中產生分割及不確定度，設計重點是單一模型內部的狀態傳播。本文未與上述方法進行同條件實測，因此僅比較其估計途徑，不宣稱性能或成本全面優於多次預測方法。

另一類作法直接從分類輸出建立分數，例如最大 softmax 機率或預測熵。最大機率可作為錯誤與分布外樣本偵測的簡單基準[20]，機率校準研究則探討輸出信心是否符合實際正確率[19]。即使一個分數能將較多錯誤排在前面，也不代表分數 0.8 就對應八成錯誤機率。因此，排序能力與數值校準必須使用不同的評估問題加以處理。本研究的不確定度 σ 由解碼器傳播的 P 經可學習讀出取得，再以停止梯度的交叉熵作為訓練訊號。這使 σ 與目前分割模型的困難程度產生關聯，但不直接區分資料噪聲及參數不確定性，也不提供完整的貝氏後驗分布。本文因而以「錯誤辨識分數」作為主要操作性定義，使用預測是否等於標註作為二元標籤，檢查 σ 是否能賦予錯誤體素較高的分數。影像梯度在本文中作為不需額外訓練的對照訊號，用以檢查 σ 是否僅具有類似強度變化的排序效果。梯度較大的位置可能是器官邊界，也可能是器官內部的影像變化，因此梯度與分割錯誤並非一對一關係。 σ 優於此對照可支持其包含超過簡單強度變化的錯誤排序資訊，但仍不能由單一對照推論它優於所有不確定度方法。

表 3 不確定度估計途徑與本文評估範圍

途徑	訊號來源	需要注意的條件
最大機率／熵 [19,20]	分類輸出分布	排序與校準需分開評估
蒙地卡羅隨機失活 [17]	同一模型的多次隨機輸出	需重複前向及一致取樣設定
深度集成 [18]	多個模型的預測差異	包含多模型訓練及推論成本
本文 σ	解碼 P 的可學習讀出	以錯誤排序為操作性定義
影像梯度對照	正規化影像的強度變化	不需額外模型訓練

2.7 深監督與資源受限的訓練

深監督將輔助損失施加於中間層，使較早的表示也收到與任務相關的梯度訊號[23]。在多尺度分割中，可於不同解碼解析度建立輔助輸出，並將標註下採樣到相同尺度進行監督。這種作法可能改善最佳化過程，但其效益會受到架構、尺度及權重設定影響。當一個方法使用深監督而另一個方法未使用時，整體分數差異便同時包含模型與訓練設計的影響。三維網路的資源限制也會影響可採用的訓練程序。Gradient checkpointing 以重算部分前向運算換取較低的中間啟動值儲存需求[24]；混合精度則以較低精度運算搭配必要的高精度數值處理，降低訓練的資源負擔[25]。本文在卷積與注意力運算中使用 bf16，自訂增益與部分對數狀態更新則維持 fp32，以減少小值、相減與對數運算的數值問題。這些工程配置的目的在於使完整模型能在單張 GPU 上穩定訓練，並不直接構成分割機制的創新。本文將硬體、輸入區塊、訓練輪數及損失配置逐項列出，讓讀者能區分架構設計與可執行條件。深監督的影響則以補充實驗呈現，與五折完整方法的結果分開分析。

2.8 評估指標與本文定位

體積重疊、表面距離與錯誤排序回答不同的評估問題。Dice 描述預測與標註區域的重疊程度，NSD 描述表面在指定距離容忍度內的一致性，AUROC 則描述連續分數對二元事件的排序能力。指標選擇必須配合研究目的與資料特性，並說明平均方式及評估範圍[26]。本文同時使用三種指標，以避免用單一平均分數概括所有功能。在多器官資料中，若直接對所有體素計算整體正確率，大器官及背景可能占據主要影響；若對器官等權平均，小器官的表現則有較清楚的呈現。本文的分割主結果排除背景，對各案例及各器官的分數進行彙總，並另列逐器官差異。對不確定度則限定預測或標註的前景聯集，使比較聚焦在器官及其誤分區

域，減少大量容易背景對排序結果的影響。綜合上述文獻，本文的研究位置是以既有 U 形 Transformer 為基礎，研究具狀態傳播的融合機制與單次前向不確定度輸出。研究問題不要求從零建立新的影像骨幹，而是檢驗結構化融合能否在既定分割流程中提供實際改善。方法章將以數學式與張量流程具體描述運算，實驗章則以整體比較與局部消融界定其有效範圍。

第三章 研究方法

3.1 整體架構與問題定義

本研究以三維腹部 CT 作為輸入，目標是在每個體素預測背景或十三種器官之一，並同時輸出一張描述分割困難程度的不確定度圖。令輸入影像為 I ，標註為 Y ，模型產生十四通道的分類 logits 與單通道的對數變異讀出 s 。分類結果由 logits 最大的類別決定，不確定度則以 $\sigma = \exp(s/2)$ 表示。分割與不確定度共用編碼及解碼表示，因此不需要為每個體素另行訓練一個獨立分類器。Swin-TK 保留 SwinUNETR 的階層式骨幹，將改動集中於編碼器的資訊更新及解碼器的多尺度融合。編碼端以卡爾曼式閘門取代原有的兩處殘差加法，並加入溫度調節與對數狀態偏置；解碼端則同時傳播特徵與正值狀態 P ，以預測、觀測及增益更新完成逐層融合。圖 2 呈現兩條路徑與輸出端的關係。圖中的狀態指網路內部張量，尺度由粗至細的變化不代表真實時間序列。

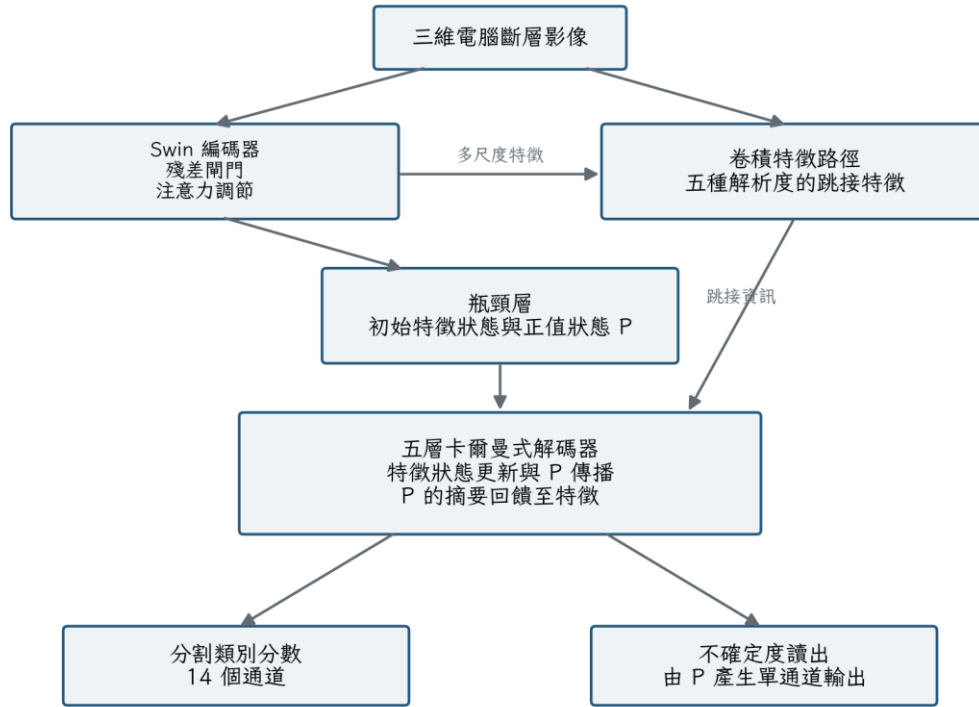


圖 2 Swin-TK 整體架構與雙輸出路徑

相較於僅在輸出端接上一個不確定度預測頭，本文使 P 在五層解碼中持續參與計算。 P 一方面控制編碼跳接與解碼預測之間的增益，另一方面經通道摘要後回饋至特徵精煉層，最後再由讀出層產生 s 。這種配置使融合過程與不確定度表示具有共同來源；是否形成有用的錯誤排序，則由第四章的 AUROC 評估，而非

由網路連接方式直接假定。模型採用基礎通道數 48，四個 Swin stage 各包含兩個區塊。以 $96 \times 96 \times 96$ 的輸入為例，Transformer 路徑由半解析度逐步降低至三十二分之一解析度，再透過五層解碼器回復原始大小。完整解析度的卷積特徵與中間尺度的編碼特徵共同構成跳接來源。表 4 列出解碼流程中的主要張量大小，批次維度省略，空間尺寸依三個方向相同的範例表示。三維 CT 經重採樣後仍可能具有不同的體積大小，表中數值因此是訓練區塊的形狀範例，而不是限制所有影像必須縮放為單一體積。完整影像推論採滑動視窗，使相同網路可處理超過訓練區塊的輸入。編碼特徵與跳接在進入同一解碼層前具有相容的空間尺寸，通道差異則由各自的可學習投影處理。

表 4 五層解碼器的張量配置

階段	空間尺寸	輸入通道	輸出通道	跳接來源
瓶頸	3^3	768	768	初始 P 由瓶頸投影
解碼第 5 層	6^3	768	384	Swin 隱藏特徵 hs[3]
解碼第 4 層	12^3	384	192	編碼特徵 enc3
解碼第 3 層	24^3	192	96	編碼特徵 enc2
解碼第 2 層	48^3	96	48	編碼特徵 enc1
解碼第 1 層	96^3	48	48	完整解析度 enc0
輸出	96^3	48	14 與 1	分類分數與 s

註：以輸入 96^3 、基礎通道 48 為例；每層 P 與該層特徵具有相同通道與空間尺寸。

3.2 具正值狀態傳播的解碼器

3.2.1 特徵預測與狀態轉移

令第 1 層的輸入特徵為 x ，對應正值狀態為 P ，較高解析度的跳接特徵為 e 。解碼器先以步長為二的三維轉置卷積形成預測特徵 x^- ，將空間解析度提高一倍； P 則以三線性插值提高解析度，再經 $1 \times 1 \times 1$ 卷積轉換通道。狀態投影與過程項各經 softplus，兩者相加形成預測狀態 P^- ，如式 (6) 所示。其中 T 表示特徵上採樣映射， U 表示空間插值， W_P 表示狀態通道投影， q 為可學習過程項的原始參數。softplus 將任意實數轉為正值，避免負值直接進入增益分母。 P^- 的兩部分分別提供來自前一尺度的資訊與本層可學習的增量；模型不需要事先指定每個器官或每個位置的雜訊大小，而是由訓練資料決定其適合的表示。

$$\begin{aligned} x^- &= T(x) \\ P^- &= \text{softplus}(W_P U(P)) + \text{softplus}(q) \end{aligned} \quad (6)$$

此處的狀態轉移不是傳統完整共變異數的 FPF^T 運算。以每個位置及通道一個正值表示，能避免建立通道平方規模的矩陣，適合三維特徵所需的記憶體條件。代價是模型不明確儲存所有通道之間的共變異關係；通道互動主要透過卷積投影與後續精煉層表達。因此， P 的核心角色是具有正值約束的融合狀態，而不是完整統計共變異數的直接估計。最深層的 P 由瓶頸特徵經 $1 \times 1 \times 1$ 卷積與 softplus 初始化。這使起始狀態依賴輸入影像，後續各尺度則以各自的參數完成轉移。若只在每層獨立產生門控權重，模型不會具有相同的跨尺度 P 路徑；本文的設計則保留前一尺度的狀態作為下一尺度更新的輸入，形成連續的融合記錄。

3.2.2 跳接觀測與增益更新

跳接特徵 e 透過兩個不同的 $3 \times 3 \times 3$ 卷積，分別產生觀測 z 及觀測雜訊狀態 R 。觀測分支負責轉換可供融合的內容， R 分支負責提供形成增益所需的正值量；兩者輸入相同，但參數與功能不同。 R 經 softplus 後加上小常數 10^{-4} ，以避免分母接近零。預測與觀測的通道數在本層一致，因此可以進行逐元素運算。式 (7) 中的所有乘除皆為逐元素運算。當 R 相對較大， K 會降低，更新較多保留 x^- ；當 P^- 相對較大， K 會提高，觀測 z 的影響增加。由於 P^- 與 R 皆為正值， K 位於零與一之間，更新可寫為 $x^+ = (1 - K)x^- + Kz$ 。這個形式使兩種來源的融合比例明確，但比例是否對應實際可靠程度，仍取決於參數經任務損失學習後的結果。

$$\begin{aligned} z &= W_z e, \quad R = \text{softplus}(W_R e) + \varepsilon \\ K &= \frac{P^-}{P^- + R} \\ x^+ &= x^- + K \odot (z - x^-), \quad P^+ = (1 - K) \odot P^- \end{aligned} \quad (7)$$

圖 3 顯示一層解碼器的資訊流程。特徵分支與狀態分支在形成 K 時交會，再分別產生更新後的特徵與 P 。此設計具有兩個互補效果：特徵差值 $z - x^-$ 提供修正方向，正值狀態提供修正幅度。模型因此不只學習如何轉換跳接內容，也學習在不同位置與通道上使用多少觀測資訊。從單層代數關係觀察，更新後 P^+ 不大於該層 P^- 。然而，下一層會重新經過通道投影與過程項累加，因此不能推論 P 從最深層到輸出一定逐層下降。相同地， K 接近零或一並不直接代表分割正確或錯誤，而是表示融合偏向哪一個來源。本文將這些內部量與最終 σ 分開定義，避免把不同尺度及不同讀出方式的數值混為同一種機率。

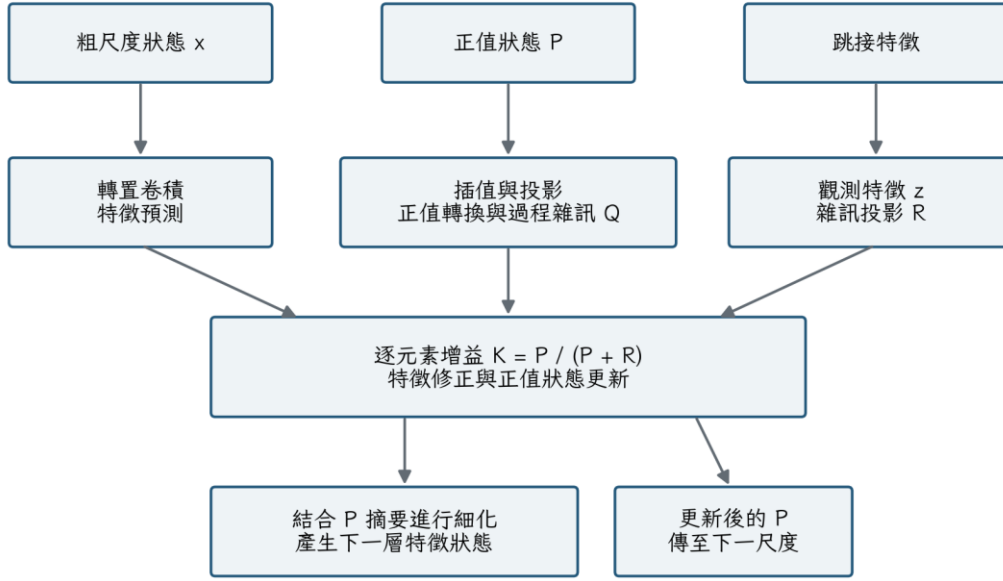


圖 3 卡爾曼式解碼器的預測、觀測與更新流程

3.2.3 狀態摘要與特徵精煉

完成增益更新後，模型先對 P^+ 沿通道取平均，得到單通道摘要，再與 x^+ 串接。串接結果通過 $3 \times 3 \times 3$ 卷積、Instance Normalization 與 Leaky ReLU，形成傳至下一層的特徵。其運算如式 (8) 所示， ϕ 表示精煉映射， C 表示通道數。這個回饋路徑讓特徵精煉層能看到目前融合狀態的摘要，而非僅接收更新後的特徵。以單通道摘要取代完整 P 串接，可以控制額外通道及卷積成本，同時保留空間分布。不過，通道平均會壓縮通道間差異，因此它是一種成本與表示能力間的選擇；本文未比較其他摘要方式，無法據此宣稱平均是最適合的聚合運算。經過精煉後的特徵已包含卷積及非線性變換，所以整層輸出不再必然位於 x^- 與 z 的逐元素區間內。凸組合的性質適用於增益更新步驟，而非整個解碼模組。這項區別也說明，Swin-TK 並未將神經網路限制為線性濾波器，而是在具有一般非線性表示能力的架構中加入可分析的融合步驟。

$$\bar{P} = \frac{1}{C} \sum_{c=1}^C P_c^+, \quad x_{\text{out}} = \phi(\text{Concat}(x^+, \bar{P})) \quad (8)$$

3.3 編碼器的卡爾曼式閘門

標準 Transformer 區塊通常將注意力或 MLP 的輸出加回原有表示[5,29]。本文在這兩個更新位置使用資料相依的增益，將先前特徵 x 與新分支輸出 y 進行混合。每個閘門具有可學習的通道參數 p ，並以線性投影 W_R 根據 y 產生觀測狀態。式 (9) 列出其定義， ε 為正值穩定常數。此處 y 是注意力或 MLP 分支的原

始輸出，並不是已經加上 x 的殘差結果。因此，更新為 $x + K(y - x)$ ，不能改寫成 $x + Ky$ 。兩種形式具有不同的保留比例：前者同時縮放舊特徵與新分支，後者則完整保留舊特徵。論文中的方法描述與實作採用前者，並以此說明閘門對特徵更新的直接作用。閘門中的 P 依通道學習， R 則隨目前 token 的分支特徵改變。這使相同通道在不同位置仍可具有不同的 K ，而不需要為每個位置儲存固定參數。 R 投影與 p 的原始參數以零初始化，經正值轉換後使初始增益接近二分之一。初始設定提供平衡的兩來源混合，隨訓練再由分割與共用表示相關梯度調整。編碼閘門的 P 與解碼器的 P 有共同的正值增益形式，但傳播方式不同。編碼閘門使用每個模組的通道參數，解碼器則將空間狀態逐層傳遞。另有一條對數狀態 ℓ 專門用於注意力偏置，下一節將說明其更新。本文保留不同符號與路徑的區別，以避免將所有帶有卡爾曼形式的元件誤認為同一個共變異數張量。

$$\begin{aligned} P &= \text{softplus}(p) + \varepsilon, \quad R = \text{softplus}(W_R y) + \varepsilon \\ K &= \frac{P}{P + R}, \quad x_{\text{new}} = x + K \odot (y - x) \end{aligned} \quad (9)$$

3.4 注意力的溫度與狀態偏置

3.4.1 查詢溫度調節

對每個查詢及注意力頭，模型由輸入特徵投影得到一個調節量，再透過 $\tanh$ 與可學習強度參數形成溫度 τ 。 τ 被限制在 0.2 至 5 的範圍內，以控制注意力分布過度尖銳或平坦的程度。若以 B 表示相對位置偏置、 M 表示位移視窗遮罩，完整注意力 logits 如式 (10) 所示。式中 i 與 j 分別表示查詢及鍵位置， h 表示注意力頭。相對位置偏置與內容相似度一起除以 τ ，對數狀態偏置則在之後加入。正值溫度主要改變同一查詢下 logits 的相對差距，不會自行改變溫度縮放前各鍵的排序；softmax 後的權重分布則會隨差距改變。鍵位置偏置具有不同作用，因為各鍵可受到不同的加減值，因而可能改變來源排序。溫度投影的權重與偏差初始化為零，使 τ 的起始值為一。調節強度使用 softplus ，因此原始強度參數為零並不同於功能關閉；初始 τ 為一來自投影輸出為零。若未來進行無溫度消融，應直接固定 $\tau=1$ ，而不是僅把原始強度設為零。這個運算定義也使消融設定可以與公式一一對應。

$$\begin{aligned} \tau_{ih} &= \text{clip}(1 + \text{softplus}(\alpha) \tanh(W_\tau X_i), 0.2, 5) \\ A_{ij}^h &= \frac{\frac{q_i^h (k_j^h)^T}{\sqrt{d}} + B_{ij}^h}{\tau_{ih}} - \mu \ell_j + M_{ij} \end{aligned} \quad (10)$$

3.4.2 對數狀態的傳播

每個輸入樣本建立單通道的對數狀態 ℓ ，起始值為零，對應指數空間中的單位正值。進入區塊後，模型先將前一狀態與可學習過程項相加，但運算在對數空間使用 logaddexp 完成，如式 (11) 所示。這樣可避免先指數化再相加所增加的數值範圍需求。注意力輸出 δ 經通道平方平均，形成每個位置的能量，再以同一樣本的空間平均值正規化。能量計算對 δ 停止梯度，並在對數運算前加入小常數。此訊號反映注意力分支輸出的相對幅度，用於更新 ℓ ；它不是由已知狀態真值計算的量測誤差，也不能直接稱為經統計驗證的創新共變異數。式 (12) 中的 b 是可學習偏移， clip 將更新後的 ℓ 限制在 -6 至 6 。更新採用 log-sigmoid 形式，使其可在對數空間中描述一個乘法式縮減。這裡的狀態傳播是特徵依賴的調節設計，目的是將前一區塊的資訊帶入後續注意力，而不是重建真實影像噪聲的物理模型。 ℓ 以鍵位置偏置 $-\mu\ell_j$ 進入式 (10)。 μ 是可學習且未限制正負的參數，因此模型可以學到不同方向的調節，不能預設其一定抑制較大的 ℓ 。對其功能的描述應以「利用狀態改變來源權重」為主；若要進一步解釋偏好方向，需要觀察訓練後的參數與特徵分布，本文目前不以未量測的方向性作為結果主張。

$$\ell^- = \text{logaddexp}(\ell, q_{\log}) \quad (11)$$

$$\begin{aligned} n &= \frac{\sum_{c=1}^C \text{stopgrad}(\delta_c)^2}{C}, \quad \hat{n} = \frac{n}{\text{mean}(n) + \varepsilon} \\ z &= \ell^- - \log(\hat{n} + \varepsilon) + b \\ \ell^+ &= \text{clip}(\ell^- + \text{logsigmoid}(-z), -6, 6) \end{aligned} \quad (12)$$

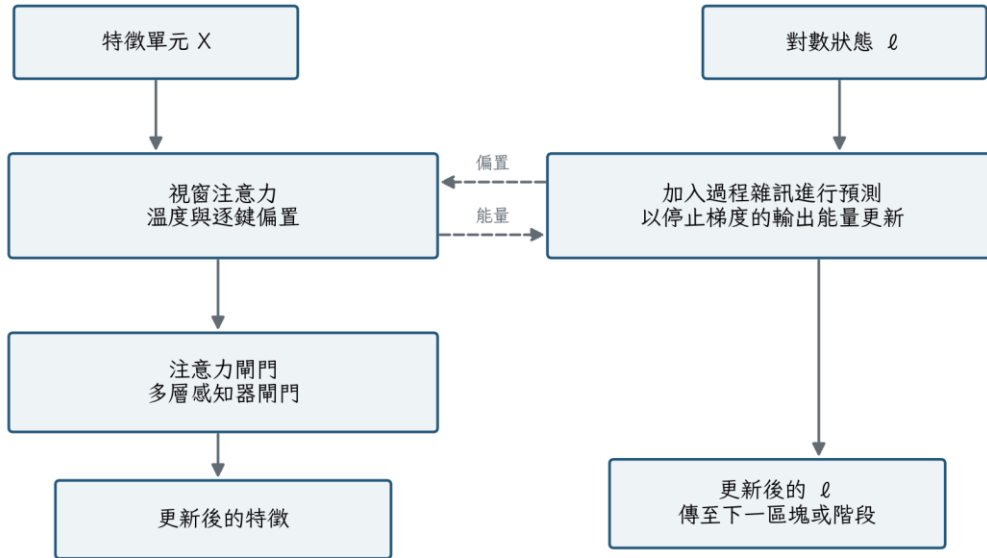


圖 4 編碼器中閘門、注意力調節與對數狀態的關係

3.4.3 空間對齊與調節作用

ℓ 與 token 特徵共用相同的空間布局，因此在視窗切分前，兩者使用一致的補邊及循環位移；進入視窗後， ℓ 依鍵位置加入對應注意力矩陣。跨 stage 下採樣時，再以插值使狀態解析度與特徵一致。這些步驟確保偏置作用於預期的來源位置，是方法內部運算的一部分。圖 5 以四個鍵的示意值比較溫度與鍵偏置。左圖使用相同 logits 改變正值溫度，可觀察分布由集中轉為平坦而排序維持；右圖則使用不同鍵偏置，使權重相對分配改變。示意數值不是由訓練模型擷取，而是用於說明兩種機制處理的不同自由度。兩者共同使用時，模型能同時調整每個查詢的分布尺度與每個來源的相對權重。在 gradient checkpointing 下，部分前向運算會在反向時重新執行。實作因此將狀態與所需資訊明確傳入對應運算，避免依賴會在重算前被覆寫的暫存變數。這項處理的目的在於維持同一區塊在前向及重算時使用相同輸入，使節省記憶體的配置不改變所定義的更新流程。

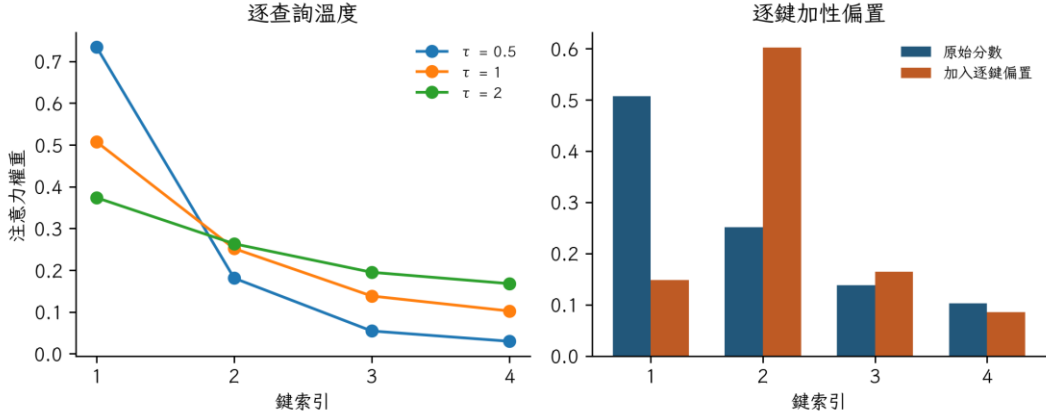


圖 5 查詢溫度與鍵位置偏置的作用示意

3.5 不確定度讀出與訓練目標

3.5.1 由最終狀態產生錯誤分數

最終解碼狀態 P_{out} 經單通道 $1 \times 1 \times 1$ 卷積產生 s ，再取 $\exp(s/2)$ 得到 σ ，如式 (13) 所示。此讀出層具有可學習權重與偏差，因此輸出不是直接將 P 的通道平均當作變異數。由於權重不限制為正，也不能假設某個 P 通道增加時 σ 必然增加；其關係由整體訓練共同決定。 P 的狀態轉移、過程項、 R 投影及 s 的讀出均包含可學習參數。將此方法稱為可學習卡爾曼式融合，強調的是增益與狀態更新的結構，而非將傳統公式直接套上固定常數。分割損失透過 K 及特徵精煉回傳至相關分支，不確定度目標則進一步約束輸出 s ，使兩者共享一部分可學習表示。

$$s = W_s P_{out} + b_s, \quad \sigma = \exp\left(\frac{s}{2}\right) \quad (13)$$

3.5.2 不確定度目標的解析性質

對體素 v ，以 CE_v 表示分割交叉熵， stopgrad 表示停止該路徑的梯度。不確定度目標採式 (14)，其中 Ω 為該訓練區塊中的體素集合。實際計算前將 s 限制於 -3 至 3 ，以控制指數權重的數值範圍。本文以此作為受異方差損失形式啟發的誤差學習代理目標，而非宣稱其完整重現分類機率的生成模型。在固定 $CE > 0$ 且暫不考慮截斷時，對單一體素的 s 微分可得式 (15)。令導數為零，最小值位於 $\log(2CE)$ ，對應 σ 的理想值為 $\sqrt{2CE}$ 。因此，在此理想條件下，較大的交叉熵會對應較大的讀出分數。若 s 達到截斷邊界，或模型讀出能力不足，實際輸出便不一定達到此點。本文 CE 項的係數為一，不能將最適解寫為 $\log(CE)$ 。Kendall 與 Gal 在常見高斯迴歸形式中包含二分之一的資料誤差係數[16]；本文使用交叉熵代理目標，係數不同，因此對 s 的定值解也不同。這個推導說明本文訓練目標的實際尺度，但並不表示 σ 已經成為經校準的錯誤機率，第四章仍以排序能力作為主要評估。停止 CE 的梯度，可以避免不確定度項沿 CE 分支直接修改分類 logits，使分類主目標保持原有權重。然而， s 依賴共用的 P 與特徵，不確定度損失仍可透過這些共用部分影響網路。因此，不能把此設計描述為與分割完全隔離；它減少的是特定梯度路徑的直接耦合，而不是取消所有共同學習效應。

$$L_{\text{unc}} = \frac{1}{|\Omega|} \sum_{v \in \Omega} [e^{-s_v} \text{stopgrad}(CE_v) + \frac{s_v}{2}] \quad (14)$$

$$\begin{aligned} \frac{\partial L}{\partial s} &= -e^{-s} CE + \frac{1}{2} = 0 \\ s^* &= \log(2CE), \quad \sigma^* = \sqrt{2CE} \end{aligned} \quad (15)$$

3.5.3 分割、深監督與總損失

主要分類輸出使用 Dice 與交叉熵之和。深監督啟用時，半、四分之一及八分之一解析度的輔助輸出分別使用 0.5、0.25、0.125 的權重；標註以最近鄰插值配合各尺度。總分割損失如式 (16)，其中 L_{DC} 表示 DiceCE， a_r 表示第 r 個輔助輸出， Y_r 表示同尺度標註。輔助輸出的目的在於提供中間層監督，推論時不需要執行這些分類頭。完整方法包含主分割、輔助分割及不確定度三部分，總目標如式 (17)。不確定度權重在前三十個 epoch 線性增加至 0.1，之後維持固定，使早期訓練先建立基本分割能力。基準模型 僅使用主輸出的 DiceCE，完整 Swin-TK 則使用上述總目標。兩者共用資料切分、訓練預算與主要最佳化設定，但損失結構並不完全相同。為分析深監督的作用，補充實驗在 Swin-TK 中關閉輔助輸出損失，保留主要分割及不確定度項。這項對照能檢查提升是否完全依賴深監督；若要再拆出不確定度訓練及各模組的個別效益，仍需額外消融。

$$L_{\text{seg}} = L_{\text{DC}}(\text{logits}, Y) + \sum_{r=1}^3 w_r L_{\text{DC}}(a_r, Y_r) \quad (16)$$

$$(w_1, w_2, w_3) = (0.5, 0.25, 0.125)$$

$$L = L_{\text{seg}} + \lambda(e) L_{\text{unc}}, \quad \lambda(e) = 0.1 \min(1, \frac{e}{100}) \quad (17)$$

3.6 訓練與推論程序

模型由隨機初始化開始訓練，未使用外部預訓練權重。每個 epoch 固定執行 250 次參數更新，總預算為 1,000 個 epoch。最佳化器採用具 Nesterov 動量的 SGD，初始學習率為 0.01，動量為 0.99，權重衰減為 3×10^{-5} 。學習率以多項式方式降低，如式 (18) 所示， e 為目前 epoch， E 為總 epoch 數。訓練使用 bfl6 自動混合精度，損失計算與需要較高數值穩定性的運算使用 fp32，並將梯度範數限制為 12。每二十五個 epoch 進行一次驗證，以驗證平均 Dice 選擇權重。模型、最佳化器與排程器狀態一併儲存，以支援中斷後續訓。表 5 將訓練與推論的主要步驟整理為可重現流程。完整體積推論使用 96^3 的滑動視窗，視窗間重疊比例為 0.5，每次處理兩個視窗。重疊區域採用推論函式的常數權重聚合 logits，再以最大類別產生分割；不使用測試時增強、模型集成或額外連通區後處理。若同時取得 s ，則將其與分類 logits 一起作為輸出通道送入滑動視窗聚合，使兩種輸出保持相同的空間位置。平均 s 後再取 $\exp(s/2)$ ，與直接平均各視窗的 σ 並不相同，因此不確定度輸出應保持固定的聚合定義。本文的程式路徑採前者；其優點是分割與對數讀出可共用一次滑動視窗呼叫。這裡的單次前向是指每個輸入視窗不需重複隨機取樣，並非完整 CT 只需要一次視窗運算。

$$\eta(e) = 0.01(1 - \frac{e}{E})^{0.9}, \quad E = 1000 \quad (18)$$

表 5 訓練與推論流程

步驟	階段	運算
1	資料準備	使用該折訓練統計轉換 CT，抽取 96^3 區塊
2	編碼	逐階段更新特徵、閘門及對數狀態 ℓ
3	初始化	由瓶頸產生特徵狀態與正值 P
4	五層解碼	依序預測、投影觀測、計算 K 並更新 x 與 P
5	讀出	產生十四類分數、單通道 s 及訓練輔助輸出
6	訓練更新	計算重疊與交叉熵損失、深監督及暖身加權不確定度損失
7	驗證	滑動視窗聚合並以平均重疊係數選擇最佳權重
8	推論	關閉輔助輸出，取得分割與 $\sigma = \exp(s/2)$

3.7 方法特性與可檢驗假說

Swin-TK 的主要設計假說是，多尺度融合若具有可學習的正值狀態及顯式增益，便可能改善對不同來源的使用方式；第二項假說是，與融合共用狀態的讀出可以學到與分割錯誤相關的排序訊號。前者由 DSC、NSD 及無深監督比較檢驗，後者由 σ 與影像梯度的 AUROC 對照檢驗。這使方法的每個主要主張都有相應的可觀察結果。就成本而言，狀態主要以逐元素張量表示，增益本身不需要矩陣反轉，但投影、觀測及精煉仍包含額外卷積。因此，成本評估必須同時考慮參數、運算量、記憶體讀寫與實際執行時間，不能只因公式簡單就推論無額外負擔。本研究已有參數量比較，實際延遲與峰值記憶體則保留為後續量測項目。本章所定義的完整方法包含多個相互作用的模組，現有結果能評估整體配置與深監督的部分差異，但不足以為每一個元件單獨分配貢獻。第四章因此先呈現整體效果，再依可用對照討論其支持範圍。這樣的實驗結構有助於將新架構的設計理由、已得到的證據與下一步驗證問題連接起來。

第四章 實驗結果與討論

4.1 資料集與交叉驗證設計

本研究使用 FLARE22Train 的五十例腹部 CT 與對應標註，器官類別包含肝臟、左右腎臟、脾臟、胰臟、主動脈、下腔靜脈、左右腎上腺、膽囊、食道、胃及十二指腸[21]。不同器官在大小、形狀與鄰近組織上具有差異，因此此資料適合觀察多尺度模型在大型實質器官、小型腺體及細長結構上的表現。本文使用所有標註類別進行共同訓練，未為特定器官建立獨立模型。五折切分先以隨機種子 42 對案例排序進行固定排列，再分成五組，每組十例。每次取其中一組作為驗證集，其餘四組共四十例作為訓練集。基準模型與 Swin-TK 使用相同的切分，因而可在相同案例上計算配對差值。五組驗證集彼此不重疊，全部合併後涵蓋五十例；圖 6 顯示此流程，橫軸為重新排列後的案例群組。

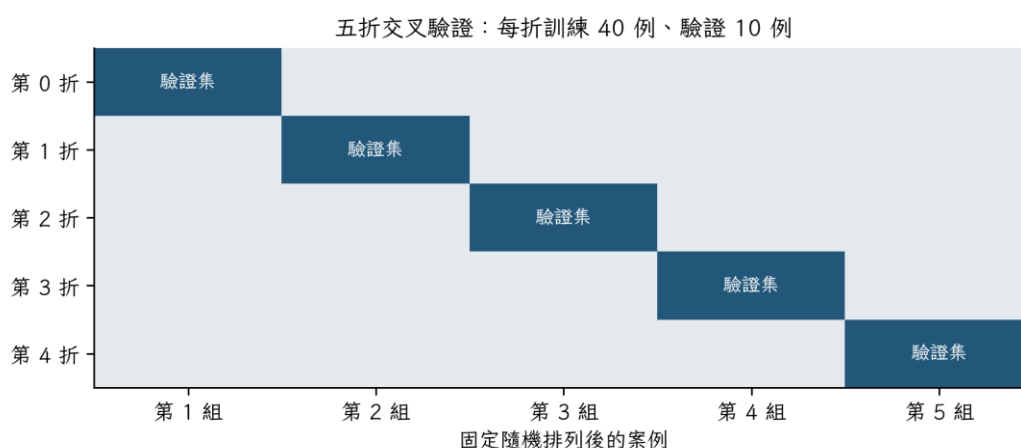


圖 6 五折交叉驗證的訓練與驗證分配

交叉驗證中的每例影像由未使用該例進行參數更新的模型預測，但該折驗證集也用於選擇最佳 epoch。因此，本文結果屬於模型選擇條件下的內部交叉驗證估計，並非完全獨立的最終測試集分數。五折模型的訓練資料亦高度重疊，不能將五十個案例配對視為五十次獨立重新訓練。這些關係會影響信賴區間與統計檢定的解讀，後文將以條件式的案例分析呈現。研究未使用 FLARE 的未標註資料，也未將公開挑戰測試集與本研究驗證集混合。不同資料使用方式、外部預訓練、模型集成及後處理可能明顯改變結果，因此本文不以官方榜單的數字直接宣稱排名。本文要回答的是：在相同的內部資料切分與主要訓練預算下，Swin-TK 相對於所實作的 SwinUNETR 是否產生可觀察的改善。

表 6 資料與交叉驗證範圍

項目	設定
資料	FLARE22Train，50 例已標註腹部 CT
輸出類別	13 種器官及背景，共 14 類
資料切分	固定種子 42，五折，每折訓練 40 例、驗證 10 例
配對比較	兩模型使用相同驗證案例
權重選擇	各折驗證平均重疊係數最佳的訓練週期
資料使用	監督式訓練；不使用未標註資料與外部預訓練
評估定位	內部交叉驗證；不等同官方獨立測試集

4.2 前處理與實驗設定

4.2.1 影像正規化及空間處理

各折的影像強度統計僅由該折訓練資料計算。程式從標註前景抽取強度樣本，取得第 0.5 與第 99.5 百分位、平均值及標準差，再將影像強度裁切至百分位區間並進行固定統計量的 z-score 正規化。此作法使驗證資料使用訓練時已決定的轉換，避免根據驗證標註重新估計強度尺度。空間方向統一為 RAS，影像以連續插值重採樣，標註則使用最近鄰插值。各折以訓練影像的間距統計設定目標解析度，第三方向皆為 2.5 mm，前兩方向約介於 0.787 至 0.820 mm。這代表 96^3 區塊在不同方向覆蓋的實際長度並不相等；三維網路使用的是重採樣後的體素鄰域，不能將體素距離直接當作各方向相同的毫米距離。完成強度與空間轉換後，使用影像前景裁切並補邊，使空間尺寸至少達到 96^3 。訓練時由器官前景及背景取樣區塊，前景與背景選擇權重設為一比一，每次影像取一個區塊。這個機率設定描述取樣中心的選擇，並不保證每個區塊中的前景體素恰好占一半。表 7 列出各折正規化參數，便於後續結果重現。

表 7 各折訓練資料的電腦斷層影像正規化參數

折別	下百分位	上百分位	平均值	標準差	間距（毫米）
0	-957	265	95.933	137.265	$0.79395 \times 0.79395 \times 2.5$
1	-957	273	97.955	136.050	$0.79395 \times 0.79395 \times 2.5$
2	-961	273	98.673	140.906	$0.78711 \times 0.78711 \times 2.5$
3	-961	266	94.700	141.915	$0.79785 \times 0.79785 \times 2.5$
4	-955	274	99.595	132.595	$0.82031 \times 0.82031 \times 2.5$

註：下、上百分位分別為前景強度的 0.5% 與 99.5%；統計不使用驗證標註。

4.2.2 資料增強與最佳化

幾何增強包含三個方向各以 0.5 機率翻轉，以及以 0.5 機率進行九十度倍數旋轉。強度增強包含尺度與偏移擾動、高斯雜訊、高斯平滑及對比調整，各操作的啟用機率為 0.15。影像與標註共同接受幾何轉換，強度操作只作用於影像。這些配置的目的在于增加訓練區塊的變化，而不改變標註類別的定義。兩種模型均使用 batch size 2、每個 epoch 250 次更新及 1,000 個訓練週期 預算，最佳化器、學習率排程與梯度裁切相同。完整 Swin-TK 額外使用第三章所述深監督及不確定度目標，基準模型使用主要輸出的 DiceCE。實驗使用單張 NVIDIA RTX 3090、24 GB 顯示記憶體進行訓練，透過混合精度與 checkpointing 配合三維網路的記憶體需求。每二十五個 epoch 的驗證使用完整影像滑動視窗，而非僅取訓練區塊計算分數。最終比較採各折驗證平均 Dice 最佳的權重。由於最佳權重的選擇依賴驗證軌跡，報告最佳值時也應觀察訓練是否穩定；本章後段呈現第 1 折的驗證曲線，作為理解模型收斂歷程的補充資料。

表 8 模型、訓練及推論的主要設定

項目	設定
基礎通道／Swin 深度	48；四個階段各 2 個區塊
訓練輸入／批次	$96 \times 96 \times 96$ ；批次大小 2
訓練預算	1,000 個訓練週期，每週期更新 250 次
最佳化器	隨機梯度下降、奈斯特羅夫動量；動量係數 0.99
學習率／權重衰減	0.01； 3×10^{-5}
排程／梯度裁切	多項式衰減，指數 0.9；範數上限 12
數值與記憶體	bf16、必要運算採 fp32、梯度檢查點
驗證頻率	每 25 個訓練週期；保留最佳平均重疊係數權重
硬體	NVIDIA RTX 3090，24 GB
完整 Swin-TK 額外損失	深監督 0.5／0.25／0.125；不確定度最大權重 0.1
推論	96^3 視窗、批次大小 2、重疊比例 0.5、常數權重聚合
測試時處理	無增強、無模型集成、無額外後處理

4.3 評估指標與統計方法

4.3.1 體積重疊與表面一致性

對一個器官，令 A 為預測體素集合， G 為標註集合，重疊相似係數（DSC）如式（19）。分數範圍為零至一，數值較高表示區域重疊較好。本文排除背景，先在案例與器官層級取得分數，再進行平均，使大型器官不會單純因體素較多而

主導全部結果。正規化表面重疊係數（NSD）以表面之間的距離是否落在器官容忍值 t 內作為判斷，概念如式（20）。 S_A 與 S_G 表示兩個表面， d 表示至另一表面的最近距離， $|\cdot|$ 表示評估實作所使用的表面量度。此指標補充 Dice 對邊界偏移的敏感度，並需明確指定距離單位與容忍值。表面評估的離散化亦可能影響結果，因此同一比較應維持相同實作[26,27]。本研究透過 MONAI 的表面 Dice 函式計算 NSD[22]，容忍值依器官分別設定，如表 9。距離使用各折重採樣後的毫米間距，結果是在重採樣空間取得。這使本文的基準模型與 Swin-TK 具有一致的內部比較條件，但不等同於已完整重現官方在原始影像空間中的最終評估流程。

$$DSC(A, G) = \frac{2|A \cap G|}{|A| + |G|} \quad (19)$$

$$NSD = \frac{|\{a \in S_A : d(a, S_G) \leq t\}| + |\{g \in S_G : d(g, S_A) \leq t\}|}{|S_A| + |S_G|} \quad (20)$$

表 9 正規化表面重疊係數的逐器官距離容忍值

標籤	器官	容忍值（毫米）
1	肝臟	5
2	右腎	3
3	脾臟	3
4	胰臟	5
5	主動脈	2
6	下腔靜脈	2
7	右腎上腺	2
8	左腎上腺	2
9	膽囊	2
10	食道	3
11	胃	5
12	十二指腸	7
13	左腎	3

註：容忍值取自實驗使用的 flare_tol.json；NSD 在各折重採樣空間計算。

4.3.2 不確定度的錯誤排序

不確定度評估將預測類別不等於標註的體素定義為正例，其餘為負例。評估範圍取預測前景與標註前景的聯集，使錯誤的背景預測與多餘器官預測皆可被納入，同時減少遠離器官的大量容易背景。 σ 與影像梯度使用相同位置及錯誤標籤，比較分數能否將錯誤體素排在正確體素之前。接收者操作特徵曲線下面積（AUROC）可由隨機抽取一個錯誤體素及一個正確體素時，前者分數較高的機

率表示；相同分數則計入一半，如式 (21)。AUROC 為 0.5 表示沒有方向性的排序能力，接近一則表示較好的錯誤排序。此指標不需先選定單一警示門檻，適合比較兩種連續訊號，但不能單獨反映固定檢查預算下的精確率。逐案例先計算 AUROC，再於同折案例間平均，最後報告五折平均與樣本標準差。影像梯度由正規化影像的三個方向差分形成幅度，用作不需訓練的影像對照。若為降低排序計算量而使用體素取樣，兩個訊號應使用相同規則；實際評估的空間範圍及取樣設定需與保存的逐折結果一起記錄，使不同版本的結果可追溯。

$$\text{AUROC} = \Pr(u_{\text{err}} > u_{\text{correct}}) + \frac{1}{2}\Pr(u_{\text{err}} = u_{\text{correct}}) \quad (21)$$

4.3.3 配對差值與不確定範圍

對每例影像，先將十三器官 Dice 平均，再計算 Swin-TK 減去 基準模型 的差值。以案例為單位進行 20,000 次有放回 bootstrap，得到平均差值的百分位區間 [30]。另外使用配對 t 檢定、Wilcoxon 符號等級檢定，以及案例層級的符號交換分析。符號交換時，同一案例的器官差值共同交換方向，以保留案例內器官之間的相關結構。統計結果用於描述目前五十例在既有權重下的配對表現，不涵蓋重新訓練、改變隨機種子及模型選擇的全部變異。本文以效果量及信賴區間為主要解讀對象，p 值作為補充。逐器官分析涉及多次比較，若未校正，單一 p 值不宜被當成某器官已獲獨立確認的證據；後文的器官分析以差值、分布與可能失敗模式為主。

4.4 五折整體分割結果

表 10 呈現主實驗。SwinUNETR 的平均 DSC 為 0.8649，Swin-TK 為 0.8755，提升 0.0106，即 1.06 個百分點；NSD 由 0.8917 提升至 0.9063，增加 1.46 個百分點。兩項指標同時提高，表示改善不只出現在體積重疊，也反映於指定容忍值下的表面一致性。圖 7 以相同零至一尺度呈現兩項指標，避免以截斷縱軸放大視覺差異。模型參數量由約 62.2 M 增至 63.6 M，增加約 1.4 M，相對增幅約 2.25%。這說明性能提升伴隨相對有限的參數增加，並非透過將骨幹規模大幅擴張取得。然而，參數量不等同於實際推論時間；新增狀態、插值及中間張量也會影響運算與記憶體。本文據此將成本主張限於已記錄的參數規模，不把它擴大為實測延遲幾乎不變的結論。

表 10 五折完整方法與基準模型的比較

方法	平均重疊係數	平均表面重疊係數	參數量 (百萬)
SwinUNETR	0.8649	0.8917	62.2
Swin-TK	0.8755	0.9063	63.6
差值	+1.06 百分點	+1.46 百分點	+1.4

註：差值單位為百分點 (pp)。DSC 採主交叉驗證彙總；NSD 採表面評估流程。完整 Swin-TK 使用深監督及不確定度損失。

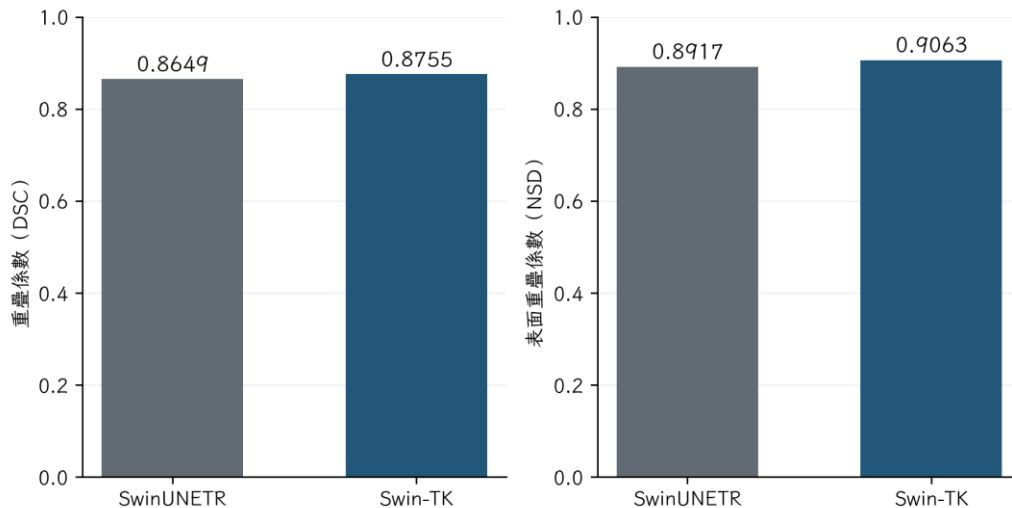


圖 7 五折主實驗的重疊與表面重疊係數比較

完整方法相對基準模型同時改動架構與部分訓練目標，所以表 10 回答的是完整 Swin-TK 配置的效果。這仍是有意義的系統比較：在相同資料與主要訓練預算下，方法達到較高的分割品質。若研究問題進一步要求將提升完全歸於某一個元件，則需觀察後續消融；本章以此區分整體方法的實用效果與個別元件的歸因。主 DSC 表採用交叉驗證彙總流程，NSD 則來自另一個同時輸出 DSC 與 NSD 的評估流程。後者的 DSC 平均為 0.8652 與 0.8758，與主 DSC 有小幅差異。兩套流程在補邊等處理與彙總細節上並非完全相同，因此本文固定主表 DSC 的來源，逐器官及同圖的 DSC/NSD 分析則使用同一份表面評估輸出，不混用其案例勝場或小數值。

4.5 逐折與逐案例分析

4.5.1 五折改善的一致性

表 11 列出五折 Dice。Swin-TK 在每一折均高於基準模型，差值介於 0.86 至 1.46 個百分點。第 0 折的基準模型分數較低，Swin-TK 在該折亦取得較大提升；其餘四折基準模型較接近，Swin-TK 仍維持正向差值。這些結果顯示整體

平均並非只由單一正向折次拉高，但各折仍具有不同的案例組成及模型訓練狀態。由五折平均方向一致，可以提高對整體趨勢的信心；但五個數值不能取代多隨機種子的重複訓練。相同資料上的不同初始化仍可能改變最佳權重與差值，尤其當比較只涉及數個小數位時，重複性需要額外量測。因此，本文將「五折皆提升」用於描述資料切分下的一致方向，不將其寫成所有訓練條件下必然提升。

表 11 主要重疊係數的逐折比較

折別	SwinUNETR	Swin-TK	差值（百分點）
0	0.8395	0.8541	+1.46
1	0.8706	0.8808	+1.02
2	0.8740	0.8838	+0.98
3	0.8704	0.8790	+0.86
4	0.8700	0.8798	+0.98
平均	0.8649	0.8755	+1.06

註：折編號沿用程式的 0-4；平均採主彙總紀錄，顯示值經四捨五入。

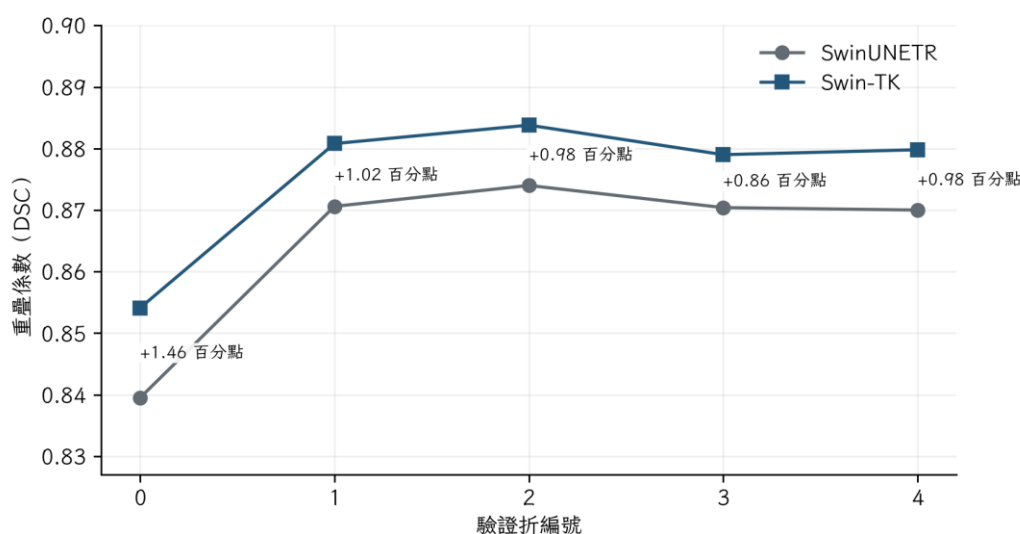


圖 8 SwinUNETR 與 Swin-TK 的逐折重疊係數

4.5.2 案例配對與效果範圍

主 DSC 彙總中，五十例有三十九例改善，占 78%；六百五十個案例與器官配對中，有四百三十九個配對改善。平均差值為 0.0106，案例 bootstrap 的 95% 區間為 0.0056 至 0.0162。以百分點表示，觀察到的平均提升為 1.06，區間為 0.56 至 1.62。這個區間完整位於正值側，支持目前案例集合中的整體正向效果。配對 t 檢定得到 $t=3.80$ 、 $p=0.0004$ ，Wilcoxon 檢定的 $W=185$ 、 $p<0.001$ ，案例層級符號交換分析的 $p=0.0001$ 。不同分析皆指向目前差值具有正向趨勢，但它們共

用同一組預測，並非彼此獨立的三次實驗。本文因此不以檢定數量增加可信程度，而是將它們視為對差值分布採用不同假設時的補充觀察。

表 12 主要重疊係數的案例配對分析

項目	結果
平均差值	+0.0106
案例自助重抽樣 95% 區間	[+0.0056, +0.0162]
正向案例	39／50
正向案例與器官配對	439／650
配對 t 檢定	$t=3.80$ ， $p=0.0004$
威爾科克森符號等級檢定	$W=185$ ， $p<0.001$
案例層級符號交換	$p=0.0001$ ，50,000 次交換

註：以現有各折最佳權重的預測為條件；區間不包含重新訓練及模型選擇的全部變異。

圖 9 使用表面評估流程中的逐案例數值，同時展示 DSC 與 NSD。對角線上方表示 Swin-TK 較高，下方表示基準模型較高；此流程中 DSC 改善為三十八例，NSD 改善為三十七例。多數點位於對角線上方，但仍有少數下降，說明平均改善並不等於每例都獲益。圖中的兩項指標與各自勝場均由同一份配對資料計算。對架構研究而言，案例分析的價值在於避免只看一個平均值。若提升僅來自少數極端案例，平均差可能缺少普遍性；目前結果則顯示多數案例改善，同時保留少量較大的負向差值。後續分析應優先檢查這些下降案例的器官範圍、影像品質及預測形狀，以判斷失敗是否具有可修正的共同來源。

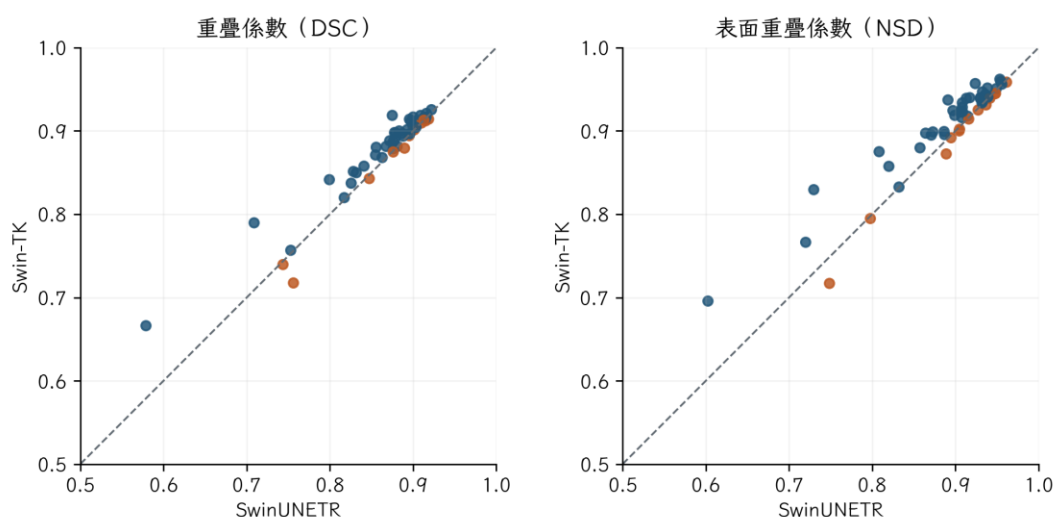


圖 9 表面評估流程中逐案例的重疊與表面重疊係數配對

4.6 逐器官表現與失敗模式

4.6.1 體積與邊界指標的共同變化

表 13 及表 14 分別列出同一表面評估流程的逐器官 DSC 與 NSD。十三個器官中有十二個器官的平均分數提高，右腎為下降項。DSC 提升較大的器官包含左腎上腺、膽囊、脾臟、胃及下腔靜脈；NSD 則以膽囊、脾臟、胃、十二指腸及下腔靜脈的改善較為明顯。圖 10 將差值以百分點呈現，便於比較兩種指標的方向。左腎上腺的 DSC 由 0.7902 提升至 0.8101，膽囊由 0.8290 提升至 0.8488，脾臟由 0.9374 提升至 0.9549。這些器官的大小與形狀不同，因此不能只以「小器官改善」概括全部結果。較合理的解讀是，完整方法在多種結構上皆呈現正向平均差值，與多尺度融合可同時利用語意及細節的設計動機相容；是否由同一模組造成，尚不能由這些平均值決定。NSD 的變化亦提供 Dice 以外的資訊。例如，膽囊增加約 3.36 個百分點，脾臟增加約 2.51 個百分點，而食道的 NSD 增加約 0.43 個百分點。這表示器官的體積重疊與容忍範圍內的表面一致性，不必然以相同比例改善。不同器官的容忍值也不相同，因此跨器官比較 NSD 時，應同時參照表 9，而不只以分數高低判斷哪個器官較容易。

表 13 表面評估流程的逐器官重疊係數

器官	SwinUNETR	Swin-TK	差值（百分點）
肝臟	0.9686	0.9781	+0.96
右腎	0.9345	0.9251	-0.94
脾臟	0.9374	0.9549	+1.75
胰臟	0.7641	0.7708	+0.67
主動脈	0.9443	0.9510	+0.67
下腔靜脈	0.8934	0.9083	+1.49
右腎上腺	0.8180	0.8251	+0.70
左腎上腺	0.7902	0.8101	+1.99
膽囊	0.8290	0.8488	+1.98
食道	0.8188	0.8308	+1.20
胃	0.8929	0.9094	+1.66
十二指腸	0.7339	0.7465	+1.26
左腎	0.9221	0.9265	+0.45

註：兩模型與各器官皆取自同一份表面評估輸出；差值由未四捨五入數值計算。

表 14 表面評估流程的逐器官表面重疊係數

器官	SwinUNETR	Swin-TK	差值 (百分點)
肝臟	0.9613	0.9768	+1.55
右腎	0.9238	0.9154	-0.85
脾臟	0.9324	0.9576	+2.51
胰臟	0.8690	0.8749	+0.60
主動脈	0.9283	0.9438	+1.55
下腔靜脈	0.8102	0.8306	+2.03
右腎上腺	0.9203	0.9263	+0.59
左腎上腺	0.8768	0.8967	+1.99
膽囊	0.7802	0.8138	+3.36
食道	0.8872	0.8915	+0.43
胃	0.9103	0.9311	+2.08
十二指腸	0.8792	0.8998	+2.06
左腎	0.9132	0.9239	+1.07

註：兩模型與各器官皆取自同一份表面評估輸出；差值由未四捨五入數值計算。

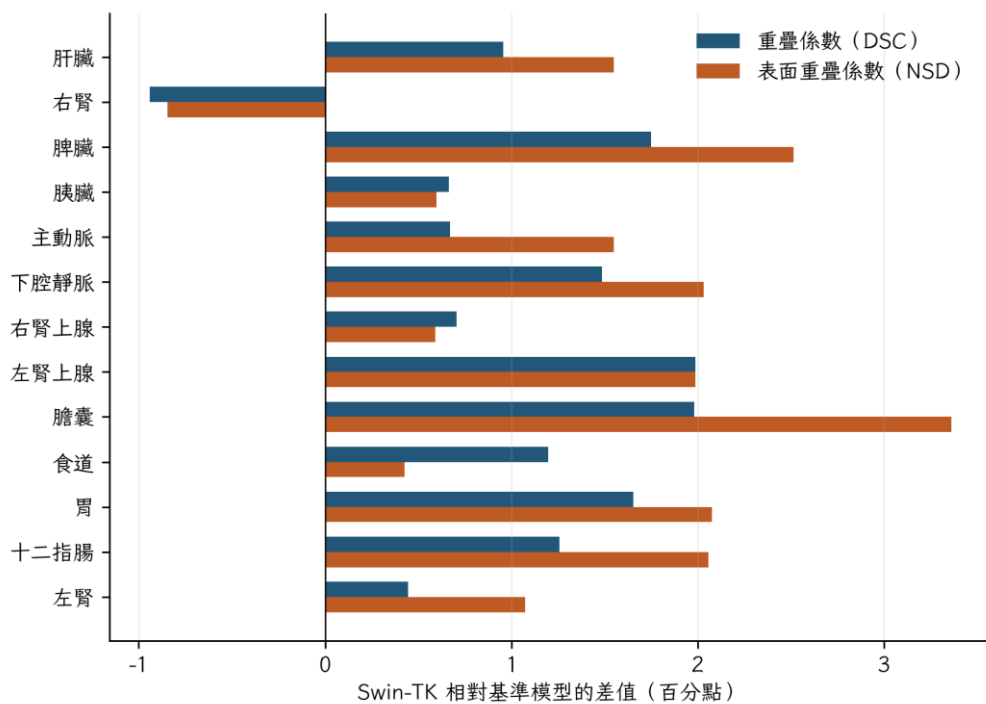


圖 10 各器官的重疊與表面重疊係數平均差值

4.6.2 右腎下降與案例差異

右腎的平均 DSC 由 0.9345 降至 0.9251，NSD 由 0.9238 降至 0.9154。此結果顯示完整 Swin-TK 並未在所有器官上全面優於 基準模型。進一步查看配對分數

可發現，少數案例具有較大的負向差值，其中一例右腎 DSC 約由 0.8055 降至 0.5407，另一例由 0.6595 降至 0.4954；這類下降會明顯影響器官平均值。目前可取得的數值不足以判定下降來自漏分、過度分割、器官辨識錯誤或前處理差異，因此本文不將某一原因寫成既定事實。較直接的後續步驟是將這些案例的 CT、標註、兩模型預測與 σ 在相同空間並列，並沿切片方向檢查錯誤範圍。若下降集中於特定視野或形狀，才有依據調整取樣、融合或後處理設計。此外，器官平均改善與大多數案例改善不是相同命題。例如，一個器官可能在許多案例小幅提高，卻被少數大幅下降抵銷。主 DSC 流程中右腎仍有三十例正向差值，但平均為負值，正好說明分布尾端的重要性。架構評估因此需要同時保留平均、勝場與失敗案例，才能判斷性能提升是否伴隨新的脆弱情境。

4.6.3 逐器官推論的界線

器官分析可協助定位後續問題，但十三次比較會增加偶然出現較小 p 值的機會。若要對個別器官提出正式推論，可使用事先定義的主要器官或多重比較控制 [28]。本文不以未校正的逐器官結果宣稱每個改善器官皆達統計顯著，而是以整體配對結果作為主要證據，逐器官結果作為效果分布與失敗分析。從方法設計看，Swin-TK 沒有針對某一器官配置專用增益或獨立參數，所有器官共用同一網路。這有利於維持統一流程，但不同器官所需的上下文與邊界資訊仍可能不同。未來若要進一步理解增益機制，可以在已定義的器官區域比較 K、P 與誤差分布，並以可重現的統計方式檢驗，而不只選取少量熱圖作說明。

4.7 深監督的補充消融

為分析完整方法的提升是否完全來自深監督，本研究在第 1 折、1,000 個訓練週期的條件下補充一組關閉深監督的 Swin-TK。表 15 的三組結果為基準模型無深監督 0.8706、Swin-TK 無深監督 0.8769，以及完整 Swin-TK 0.8808。圖 11 將三組設定依比較順序排列。當兩模型皆未使用深監督時，Swin-TK 仍提高 0.0063，即 0.63 個百分點。這項結果支持目前方法的分割改善並非完全依賴中間層監督，對新架構的研究價值具有正向意義。另一方面，Swin-TK 加入深監督後再提高 0.0039，即 0.39 個百分點，表示輔助監督在此設定下也提供額外效益。若僅沿此比較順序將總差值 1.02 個百分點拆開，0.63 與 0.39 分別約占 61.8% 與 38.2%。這是兩段觀察差值的比例，並非固定的因果貢獻。因為目前沒有基準模型加入深監督的第四個條件，也沒有多折及多種子重複，模型與深監督之間的交互作用尚未完整分離。因此，本文的主要表述採「無深監督時仍提升 0.63 個百分點」，不以「六成必然來自卡爾曼機制」作為結論。

表 15 第 1 折深監督補充比較

方法	深監督	重疊係數	相鄰差值（百分點）
SwinUNETR	否	0.8706	—
Swin-TK	否	0.8769	+0.63
Swin-TK	是	0.8808	+0.39

註：三組皆為 1,000 個訓練週期；本表為單折觀察，未提供重複訓練誤差範圍。

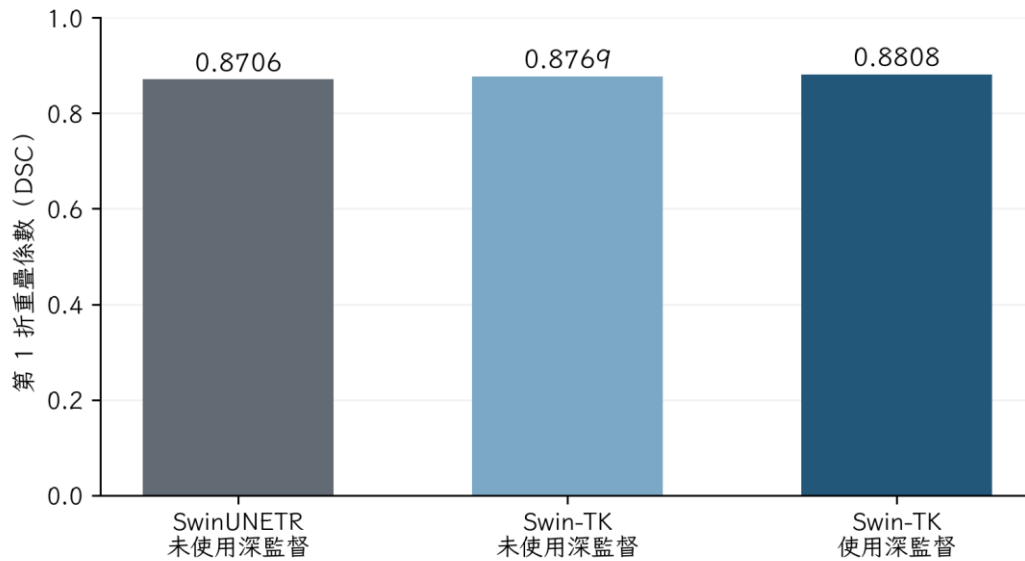


圖 11 第 1 折的深監督補充消融結果

無深監督的 Swin-TK 仍保留不確定度目標及全部架構改動，所以此比較支持的是該配置在沒有深監督下的有效性，而非純粹某一個解碼公式的孤立效益。若要完成更細的歸因，可以逐步比較僅解碼器、僅編碼器、固定門控、無溫度、無對數狀態及無不確定度損失等設定。這些設定需要各自完成訓練後才可列為結果，目前不以未完成實驗補填分數。這組消融補充了原先主實驗無法回答的問題。完整五折結果建立方法效益，單折無深監督結果則顯示該折提升並非全由深監督提供。結合兩者，可以同時描述完整配置的分割效果與其中一項訓練因素的影響，並據此規劃後續模組分析。

4.8 不確定度的錯誤辨識能力

表 16 列出五折 σ 與影像梯度的 AUROC。 σ 的平均為 0.7764，樣本標準差為 0.0387；影像梯度的平均為 0.6094，樣本標準差為 0.0411。以三位小數表示，分別為 0.776 ± 0.039 與 0.609 ± 0.041 。兩者平均差約為 0.1670，且五折皆由 σ 取得較高分數。 σ 的逐折值介於 0.7242 至 0.8178，梯度介於 0.5651 至 0.6537。兩者跨

折波動都不可忽略，但 σ 在每一折均維持優勢，差值約介於 0.1477 至 0.1840。這表示在目前評估中，學得的不確定度分數能比簡單影像強度變化更有效地排序錯誤體素，並非只在一個特定折次出現改善。影像梯度本來就可能對器官邊界產生較高反應，而分割錯誤也常涉及邊界，因此它是合理的低成本對照。 σ 優於梯度，支持模型狀態與訓練目標提供了額外的任務相關資訊。不過，此結果尚未比較 softmax 熵、最大類別機率、重複隨機預測或模型集成，故不能延伸為 σ 優於所有不確定度估計方法。

表 16 不確定度與影像梯度的逐折錯誤辨識能力

折別	不確定度 σ	影像梯度 $ \nabla I $	差值
0	0.7242	0.5664	0.1578
1	0.8178	0.6361	0.1817
2	0.7894	0.6255	0.1639
3	0.8014	0.6537	0.1477
4	0.7491	0.5651	0.1840
平均 $\pm$ 標準差	0.776 ± 0.039	0.609 ± 0.041	0.1670

註：標準差為五折樣本標準差；差值平均採未四捨五入的逐折數值計算。

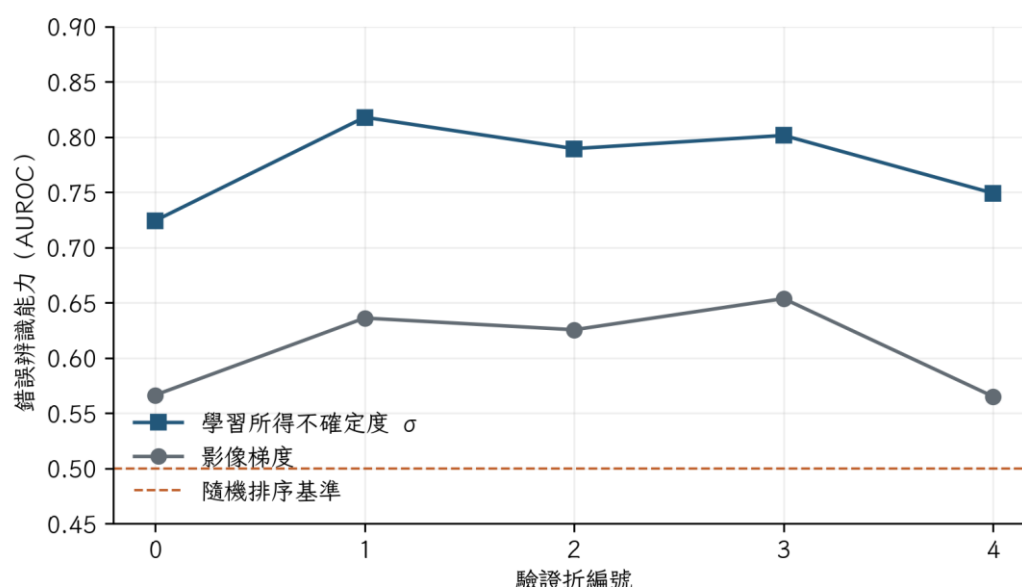


圖 12 不確定度與影像梯度的逐折錯誤辨識能力

AUROC 描述的是排序能力，不是 σ 數值的機率意義。當 $\sigma=1$ 時，不能解讀為某個固定錯誤率，也不能直接設定臨床警示門檻。若後續應用需要標示最應優先檢查的區域，可以另外以固定比例的高分體素評估錯誤召回率與誤報量；若需要機率解釋，則應補充校準程序及相應指標。這些工作建立在目前的正向排序證

據之上，與分割主結果具有不同的評估目的。由於同一器官內相鄰體素具有空間相關性，不能把數十萬體素直接當成等量獨立樣本來宣稱極高統計把握度。本研究以案例及折次進行彙總，五折標準差描述不同驗證群組的波動。若要建立 AUROC 差值的信賴區間，適合以案例為重取樣單位，保留同一案例內兩個訊號的配對關係，而不是逐體素任意重抽。

4.9 收斂行為與視覺化觀察

圖 13 由第 1 折的訓練紀錄繪製，呈現每次驗證的平均 Dice。兩種模型均在早期快速提高，後續逐漸進入較緩慢的改善區間。Swin-TK 在後期取得較高的最佳驗證結果，與表 11 的第 1 折數值一致。曲線反映完整配置的訓練歷程，不能單獨將早期或後期差異歸因於某個元件。最佳值與最後一個 epoch 的值並不必然相同，因此本文固定採驗證最佳權重作比較。訓練預算相同有助於控制更新次數，但兩模型各步運算仍不同，故不能只由 epoch 數推算相同牆鐘時間。若後續要比較效率，應以相同硬體記錄每步時間、完整體積推論延遲及峰值記憶體，並保留暖機與同步條件。

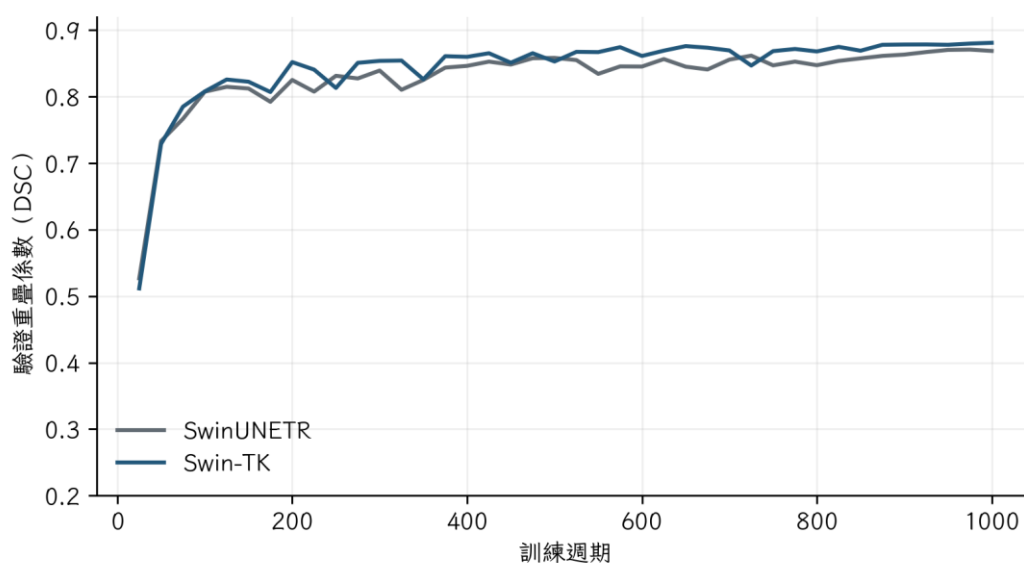


圖 13 第 1 折訓練期間的驗證重疊係數曲線

為說明分割與不確定度的空間關係，圖 14 與圖 15 使用研究開發階段保存的同一切片示例。這些圖片用於展示輸出形式與讀圖方式，並不作為上述五折版本的新增定量證據。圖 14 依序呈現 CT、預測、標註及不確定度，可觀察大範圍器官輪廓與局部邊界差異；不確定度在部分接觸面及結構附近較高，但高值區也包含正確分割位置。圖 15 左側以紅色標示錯誤位置，右側以青色呈現該圖選出的最高 5% 不確定度位置。兩者在部分區域有重疊，也有明顯不一致，說明不確定

度應被視為需要量化檢驗的連續分數。單一切片的視覺相似性無法取代三維案例與逐折 AUROC，亦無法單獨推論高分區域的錯誤比例。實際分析中，視覺化最有價值的用途是提出可檢驗問題。例如，若高不確定度集中在正確器官邊界，可能需要檢查排序是否主要反映邊界；若大面積錯誤沒有高分反應，則需要分析讀出是否漏掉特定失敗型態。本文以影像梯度對照及逐折 AUROC 回答第一部分問題，第二部分則需要在已確認的最終模型預測上建立更完整的失敗案例整理。

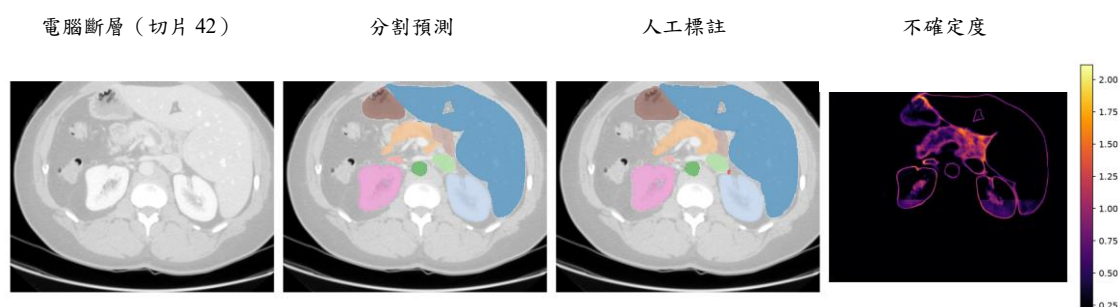


圖 14 開發階段示例的電腦斷層影像、分割、標註與不確定度

分割錯誤位置

不確定度最高的 5% 區域

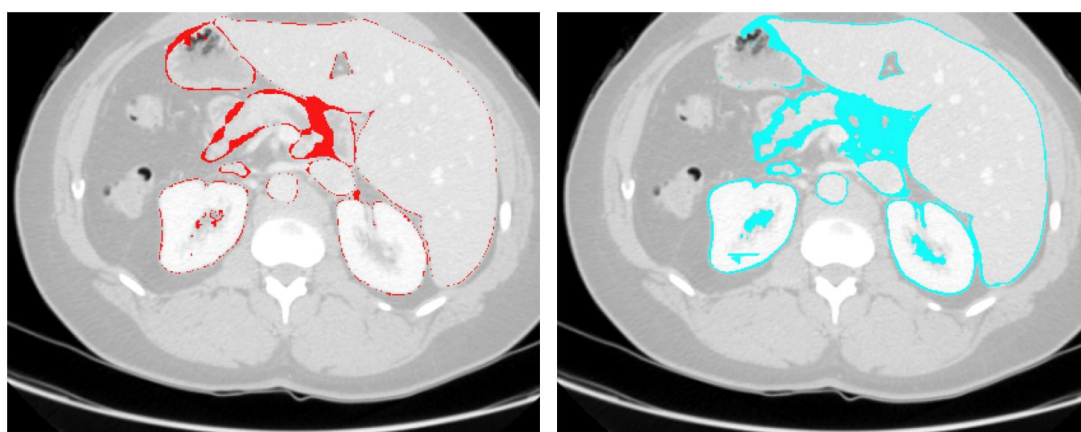


圖 15 開發階段示例的錯誤位置與高不確定度區域

4.10 綜合討論與研究限制

本研究的結果形成三層互相補充的證據。第一，完整 Swin-TK 在五折 DSC 與 NSD 上均高於基準模型，且五折 Dice 差值皆為正。第二，第 1 折關閉深監督後仍有 0.63 個百分點的提升，支持方法效益不完全依賴輔助監督。第三， σ 的 AUROC 在五折皆高於影像梯度，顯示模型除分割輸出外，也提供可辨識部分錯誤位置的訊號。表 17 將主張與證據範圍並列。就新架構的貢獻而言，本文的價值在於提出可明確實作與分析的資訊融合方式，並以實驗展示其在既有骨幹上的效果。卡爾曼式增益、跨尺度 P 與編碼器調節不是由分數自動證明的理論最優解，但它們形成具一致更新概念的模型設計。現有正向結果支持繼續研究此方

向，也使論文具備問題、方法、實作、驗證及討論的完整鏈結。目前最主要的限制是資料規模與實驗覆蓋。五十例內部交叉驗證不足以涵蓋不同醫院、掃描條件及特殊解剖變化；單折深監督消融也不能量測所有訓練隨機性的影響。此外，完整方法同時包含多個元件，尚未有完整的因子式比較。因此，後續工作宜優先增加對核心機制有辨識力的對照，而不是只增加更多相近的總分表格。

表 17 主要研究主張與目前證據

研究主張	支持結果	解讀範圍
完整方法改善分割	五折重疊係數 +1.06、表面重疊係數 +1.46 百分點	目前內部資料及訓練配置
提升不完全依賴深監督	無深監督時重疊係數 +0.63 百分點	第 1 折；保留不確定度目標
σ 具錯誤排序資訊	五折錯誤辨識能力高於梯度	不是機率校準結論
參數增加相對有限	6,220 萬 $\rightarrow$ 6,360 萬	尚未代表實測延遲結果

第二項限制涉及不確定度的語意與評估。 σ 由可學習狀態讀出，受代理損失訓練，現有證據支持的是錯誤排序，並未證明其為真實物理噪聲、完整後驗變異或已校準機率。這不否定其作為分析訊號的價值，而是界定應採用的功能名稱與驗證方式。未來加入不同簡單基準、選擇性風險曲線與跨域資料，能進一步檢查其適用性。第三項限制是評估流程的一致化。本文已分清主 DSC 彙總與表面評估的來源，但後續仍宜使用統一的前處理、輸出還原及指標程式，並保存案例識別資訊與模型版本。體積重疊主要由離散標籤決定，表面距離還受空間間距與邊界離散化影響；即使採用相同器官容忍值，若評估空間不同，結果仍不適合直接與外部數字比較。因此，統一流程的目的不只是減少小數點差異，也在於建立同一套清楚的比較條件。在保存結果時，每例應將預測標籤、對應空間資訊與不確定度輸出綁定至同一權重，並以同一案例清單計算各項指標。這可以讓右腎下降、案例差值與不確定度反應在相同資料上交叉分析，避免只保留平均值而失去追查來源的能力。後續新增消融亦應沿用此流程，將訓練設定與完成狀態一併記錄。未完成的實驗在完成前不列入結論，使目前證據與後續擴充保持可追溯的關係。

第五章 結論與未來工作

5.1 研究結論

本研究針對三維腹部多器官分割中的多尺度融合問題，提出 Swin-TransKalman。方法以 SwinUNETR 為基礎，將卡爾曼式預測與更新引入解碼器，使粗尺度特徵、跳接觀測及正值狀態共同決定逐元素融合增益。更新後的 P 沿解析度逐層傳播，並透過通道摘要參與特徵精煉，最後由可學習讀出產生不確定度圖。此設計將融合與不確定度建立在共用狀態上，形成可端到端訓練的三維分割模型。編碼器方面，本文使用資料相依的閘門混合既有特徵與注意力、MLP 分支輸出，並以查詢溫度及對數狀態偏置調節注意力。這些元件分別控制更新比例、權重分布的集中程度及不同來源的相對影響，與解碼器共同構成完整 Swin-TK。模型保留既有階層式骨幹的多尺度能力，將創新集中於資訊傳遞方式，因此能在相對有限的參數增加下進行研究與比較。

在 FLARE22Train 五十例的五折內部交叉驗證中，完整 Swin-TK 的平均 DSC 為 0.8755，高於 SwinUNETR 的 0.8649；NSD 為 0.9063，高於基準模型的 0.8917，分別提升 1.06 及 1.46 個百分點。五折 Dice 差值皆為正，主 DSC 彙總中三十九例改善，平均差值的案例 bootstrap 95% 區間為 0.0056 至 0.0162。這些結果支持所提出完整方法在本研究條件下具有分割改善。補充深監督消融顯示，在第 1 折、1,000 個訓練週期且兩模型皆未使用深監督時，Swin-TK 的 Dice 仍由基準模型的 0.8706 提高至 0.8769，增加 0.63 個百分點；加入深監督後再提升至 0.8808。此結果支持改善並非完全來自深監督，也說明完整方法可從輔助監督獲得額外收益。由於目前為單折比較，本文不將兩段差值視為各元件的固定因果貢獻。不確定度評估中， σ 的五折 AUROC 為 0.776 ± 0.039 ，影像梯度為 0.609 ± 0.041 ，五折皆由 σ 取得較高結果。這顯示模型學得的分數包含超過簡單強度變化的錯誤排序資訊。本文因而完成兩項互補目標：提升既定分割流程的平均品質，以及在單一模型內提供與部分錯誤位置相關的空間訊號。

5.2 研究貢獻與適用範圍

本文的貢獻在於將跨尺度狀態、編碼器調節與不確定度讀出整合成可實作的三維分割方法，並建立相應的實驗證據。明確的運算定義使後續研究能沿相同介面更換個別元件，進一步檢查結構化融合與一般門控之間的差異。方法的適用範圍以目前資料與訓練設定為限。整體平均改善不代表所有器官與案例皆受益，右腎的負向平均差值即提供重要反例。 σ 也應被解讀為由任務訓練得到的錯誤排序分數，尚未具有可直接對應實際錯誤率的校準保證。維持這些界定，有助於後續工作針對真正未解的問題設計實驗，而不必先建立過大的效能承諾。

5.3 未來工作

後續最優先的工作是擴充具有辨識力的消融。可在相同資料切分與訓練預算下，完成基準模型與 Swin-TK 各自有、無深監督的四組配置，以分析兩者交互作用；再比較僅解碼器、僅編碼器及完整模型，進一步拆分狀態傳播、門控、溫度與偏置的作用。對重要差值增加多個隨機種子，則可估計初始化與取樣造成的波動，讓目前單折結果獲得更完整的重複性證據。第二個方向是強化不確定度的功能驗證。首先加入最大 softmax 機率及預測熵等容易重現的基準，再以案例為單位估計 AUROC 差值區間。若應用目標是協助人工檢查，可評估固定檢查比例下找回多少錯誤，以及保留較低不確定度區域後的分割風險。這些指標能將抽象排序能力轉換為更具體的使用情境，同時避免直接把未校準分數當作機率。

第三個方向是研究跨資料來源與失敗情境。可在其他腹部資料、不同掃描條件或獨立保留集上檢查分割及不確定度是否維持效果，並針對右腎等下降案例進行一致空間的可視化與錯誤分類。若能確認特定形狀、視野或影像品質與失敗的關係，便可有依據地調整區塊取樣、融合方式或訓練目標，而不是僅增加模型容量。最後，應完成統一評估與成本量測。將所有分割輸出還原至明確定義的評估空間，使用同一套案例索引與指標程式，可使 DSC、NSD、視覺化及不確定度保持一致。配合相同硬體下的推論延遲與峰值記憶體紀錄，則能補足參數量以外的實務成本。這些延伸可使 Swin-TK 從目前具正向實證的架構研究，逐步發展為更完整、可比較的三維分割方法。

参考文献

- [1] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional networks for biomedical image segmentation,” in Proc. MICCAI, 2015, pp. 234–241. doi: 10.1007/978-3-319-24574-4_28
- [2] Ö. Çiçek et al., “3D U-Net: Learning dense volumetric segmentation from sparse annotation,” in Proc. MICCAI, 2016, pp. 424–432. doi: 10.1007/978-3-319-46723-8_49
- [3] F. Milletari, N. Navab, and S.-A. Ahmadi, “V-Net: Fully convolutional neural networks for volumetric medical image segmentation,” in Proc. IEEE 3DV, 2016, pp. 565–571. doi: 10.1109/3DV.2016.79
- [4] F. Isensee et al., “nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation,” *Nature Methods*, vol. 18, no. 2, pp. 203–211, 2021. doi: 10.1038/s41592-020-01008-z
- [5] A. Vaswani et al., “Attention is all you need,” in *Advances in Neural Information Processing Systems*, vol. 30, 2017. [Online]. Available: NeurIPS proceedings.
- [6] A. Dosovitskiy et al., “An image is worth 16×16 words: Transformers for image recognition at scale,” in Proc. ICLR, 2021. <https://openreview.net/forum?id=YicbFdNTTy>
- [7] Z. Liu et al., “Swin Transformer: Hierarchical vision transformer using shifted windows,” in Proc. IEEE/CVF ICCV, 2021, pp. 9992–10002. doi: 10.1109/ICCV48922.2021.00986
- [8] A. Hatamizadeh et al., “UNETR: Transformers for 3D medical image segmentation,” in Proc. IEEE/CVF WACV, 2022, pp. 1748–1758. doi: 10.1109/WACV51458.2022.00181
- [9] Y. Tang et al., “Self-supervised pre-training of Swin Transformers for 3D medical image analysis,” in Proc. IEEE/CVF CVPR, 2022, pp. 20698–20708. doi: 10.1109/CVPR52688.2022.02007
- [10] J. Chen et al., “TransUNet: Rethinking the U-Net architecture design for medical image segmentation through the lens of transformers,” *Medical Image Analysis*, vol. 97, Art. no. 103280, 2024. doi: 10.1016/j.media.2024.103280
- [11] H.-Y. Zhou et al., “nnFormer: Volumetric medical image segmentation via a 3D transformer,” *IEEE Trans. Image Process.*, vol. 32, pp. 4036–4045, 2023. doi: 10.1109/TIP.2023.3293771
- [12] O. Oktay et al., “Attention U-Net: Learning where to look for the pancreas,” in Proc. Medical Imaging with Deep Learning (MIDL), 2018. <https://openreview.net/forum?id=Skft7cijM>
- [13] J. Ma, F. Li, and B. Wang, “U-Mamba: Enhancing Long-range Dependency for Biomedical Image Segmentation,” *arXiv:2401.04722*, 2024. <https://arxiv.org/abs/2401.04722>
- [14] R. E. Kalman, “A New Approach to Linear Filtering and Prediction Problems,” *Journal of Basic Engineering*, vol. 82, no. 1, pp. 35–45, 1960. doi: 10.1115/1.3662552
- [15] G. Revach et al., “KalmanNet: Neural network aided Kalman filtering for partially known dynamics,” *IEEE Trans. Signal Process.*, vol. 70, pp. 1532–1547, 2022. doi: 10.1109/TSP.2022.3158588

- [16] A. Kendall and Y. Gal, “What uncertainties do we need in Bayesian deep learning for computer vision?,” in *Advances in Neural Information Processing Systems*, vol. 30, 2017. [Online]. Available: [NeurIPS proceedings](#).
- [17] Y. Gal and Z. Ghahramani, “Dropout as a Bayesian approximation: Representing model uncertainty in deep learning,” in *Proc. ICML, PMLR*, vol. 48, 2016, pp. 1050–1059. <https://proceedings.mlr.press/v48/gal16.html>
- [18] B. Lakshminarayanan, A. Pritzel, and C. Blundell, “Simple and scalable predictive uncertainty estimation using deep ensembles,” in *Advances in Neural Information Processing Systems*, vol. 30, 2017. [Online]. Available: [NeurIPS proceedings](#).
- [19] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, “On calibration of modern neural networks,” in *Proc. ICML, PMLR*, vol. 70, 2017, pp. 1321–1330. <https://proceedings.mlr.press/v70/guo17a.html>
- [20] D. Hendrycks and K. Gimpel, “A baseline for detecting misclassified and out-of-distribution examples in neural networks,” in *Proc. ICLR*, 2017. <https://openreview.net/forum?id=Hkg4TI9xl>
- [21] J. Ma et al., “FLARE Challenge,” official project repository. Accessed September 20, 2026. <https://github.com/JunMa11/FLARE>
- [22] M. J. Cardoso et al., “MONAI: An open-source framework for deep learning in healthcare,” *arXiv:2211.02701*, 2022. <https://arxiv.org/abs/2211.02701>
- [23] C.-Y. Lee et al., “Deeply-supervised nets,” in *Proc. AISTATS, PMLR*, vol. 38, 2015, pp. 562–570. <https://proceedings.mlr.press/v38/lee15a.html>
- [24] T. Chen et al., “Training Deep Nets with Sublinear Memory Cost,” *arXiv:1604.06174*, 2016. <https://arxiv.org/abs/1604.06174>
- [25] P. Micikevicius et al., “Mixed precision training,” in *Proc. ICLR*, 2018. <https://openreview.net/forum?id=r1gs9JgRZ>
- [26] L. Maier-Hein et al., “Metrics reloaded: Recommendations for image analysis validation,” *Nature Methods*, vol. 21, no. 2, pp. 195–212, 2024. doi: 10.1038/s41592-023-02151-z
- [27] Google DeepMind, “Surface distance metrics,” official implementation. Accessed September 20, 2026. <https://github.com/google-deeppmind/surface-distance>
- [28] Y. Benjamini and Y. Hochberg, “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing,” *Journal of the Royal Statistical Society: Series B*, vol. 57, no. 1, pp. 289–300, 1995. doi: 10.1111/j.2517-6161.1995.tb02031.x
- [29] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE CVPR*, 2016, pp. 770–778. doi: 10.1109/CVPR.2016.90
- [30] B. Efron, “Bootstrap Methods: Another Look at the Jackknife,” *The Annals of Statistics*, vol. 7, no. 1, pp. 1–26, 1979. doi: 10.1214/aos/1176344552