航測及遙測學刊　第三十一卷　第 1 期　第 47- 62 頁　民國 115 年 3 月
J Journal of Photogrammetry and Remote Sensing
Volume 31, No.1, 2026, pp. 47 – 62
DOI：10.6574/JPRS.202603_31(1).0004

應用深度資訊於混凝土橋梁結構物影像拼接之研究

林于婷 [1*]　高書屏 [2]　王豐良 [3]　林志憲 [4]

摘要

影像拼接可擴展視野、消除盲區，但場景深度差異容易導致視差與重影。為此，本研究整合單影像深度估計與語義分割模型，建立橋梁立面影像拼接流程，重建完整結構外觀圖作為損壞分析和管理底圖。透過遷移學習建置橋側影像數據集，沿用預訓練參數訓練 RGB-D 語義分割模型，mIoU 達 86.44%、mAcc 91.24%、召回率 92.11%、F1-score 91.56%，展現穩定性與泛化能力，並藉其成果間接驗證深度估計模型準確性。針對影像傾斜導致的幾何錯位，利用深度圖重建點雲校正。拼接精度比較顯示，結合分割模型與校正影像之平均 SSIM 為 0.6807 高於傳統方法 0.5081，證實本研究方法在精度與視覺一致性上的優勢。

關鍵詞：橋梁立面檢測、影像拼接、深度估計、語義分割、影像幾何校正

1. 前言

臺灣地形多山谷、河川縱橫，陸路交通高度依賴橋梁，使其成為重要交通設施。混凝土橋梁長期受氣候與載重影響易老化，需定期巡檢以確保安全。目前檢測多仰賴目視與簡易工具，檢測人員需搭乘作業車近距離觀察，過程耗時且具主觀性與檢測盲區。近年無人機(UAV)影像已成為重要輔助，可多角度進入難以接近區域，具便利、高效與低成本等優勢，提升檢測全面性與精準度。然而，單張影像僅呈現局部外觀，難以全面掌握橋梁狀態，需透過影像拼接重建完整結構外觀圖。然而，當影像場景深度差異顯著時，拼接結果容易在深度方向產生幾何錯位，進而出現偽影或拉伸現象，尤其在幾何結構複雜的混凝土橋梁中更難完全消除。解決此類問題的關鍵在於提升拼接過程中的特徵點檢測與對齊能力，以克服視差造成的錯位，確保拼接成果的幾何一致性與視覺品質。

無人機是目前被廣泛應用的技術之一，在建築監測與檢查中備受矚目。然而，在災後評估與建築管理中，準確偵測建物立面的損毀情形與類型仍是關鍵課題，對提升分類準確度與決策支援具有顯著影響(Tu *et al.*, 2017)。為提升損壞判讀的準確性與效率，目前多數研究以二維影像或三維點雲為基礎，結合深度學習方法進行建築物缺陷識別與結構損壞判斷，支援自動化的立面目視檢查與建築狀況評估。影像拼接則可透過整合多張影像所獲得的視覺資訊，產生視覺感知更完整的合成影像，提供比單一影像更豐富的場景描述。影像對齊是拼接演算法的核心，即便使用低成本的商用相機，也能產生自然連續的全景影像。(Kekec *et al.*, 2014)

基於特徵匹配的影像拼接透過兩張影像中對應的特徵點，計算用於投影變形的單應性矩陣(Homography Matrix)，以完成影像對齊與融合。其中，Lowe (2004)提出的 SIFT(尺度不變特徵轉換)為經典演算法，其描述子在現有局部描述子中表現優異(Liao *et al.*, 2013)，具備良好性能與穩健性，因而在眾多影像拼接演算法中廣受採用(Tang *et al.*, 2023)。後續亦有多項改進方法被提出，主要著重於提升計算效率(Ma *et al.*, 2016)。

[1] 國立中興大學土木工程學系 碩士
[2] 國立中興大學土木工程學系 教授
[3] 健行科技大學應用空間資訊系 助理教授
[4] 國立中興大學土木工程學系 博士
[*] 通訊作者，電話: 0919-907-056, E-mail: venusborn2023@gmail.com

收到日期：民國 114 年 09 月 03 日
修改日期：民國 114 年 11 月 21 日
接受日期：民國 115 年 02 月 11 日

在大視差場景下，非重疊區難以兼顧對齊精度與整體變形控制(Li *et al.*, 2021)。透過全域單應性預扭曲，再以能量函數優化對齊，視為網格扭曲問題，能更有效處理視差(Xiang *et al.*, 2018)。Zaragoza *et al.* (2013) 提出的 As-Projective-As-Possible(APAP) 演算法，通過估計投影扭曲並將圖像劃分為網格，逐一估計每個網格的變換。然而，網格變形方法雖能比傳統全局變換更有效地處理視差問題，但在低紋理影像中容易因特徵不足與辨識性低導致匹配不穩與變換誤差，產生明顯錯位(Xiang *et al.*, 2018)。此外，若影像重疊區涵蓋多個幾何平面，僅依賴全域單應性推導出單一全域相似性變換並不足夠(Lin *et al.*, 2015)。

不同於網格變形方法，場景分割可為每個像素賦予語義標籤，將影像劃分為具語義意義的區域。Cai *et al.* (2023) 結合語義分割辨識多幅影像中對應的平面區域，建立平面間的變換與高度關係以完成正射影像拼接；Kao *et al.* (2024) 則先辨識裂縫區域作為 ROI (Region of Interest)，僅對 ROI 偵測 SIFT 關鍵點以降低匹配數與運算時間，並以原始影像補足紋理資訊確保拼接精度。然而，在幾何結構複雜或紋理相近的場景中，僅依賴 RGB 影像進分割效果有限，故學者們提出結合深度資訊的 RGB-D 語義分割模型 (Wang *et al.*, 2021)。Yin *et al.* (2023)基於 RGB-D 預訓練框架提出 DFormer 模型，透過 AdaBins (Bhat e*t al.*, 2021)生成大規模對應的 RGB 與深度影像對進行多模態訓練，於編碼階段強化 RGB 與深度特徵交互，解碼階段則聚合來自編碼器最後三個層級的多尺度特徵。在多組 RGB-D 分割與檢測資料集上取得領先準確率。因此，本研究導入 RGB-D 語義分割技術引導特徵對應，以提升影像拼接的匹配精度與品質。

單目深度估計僅需單台攝影機即可恢復影像深度資訊，應用更為廣泛(Khan *et al.*, 2020)。深度學習方法在單目深度估計中取得了最佳結果，其中結合語義分割更進一步提升準確性(Rajapaksha *et al.*, 2024)。Bhat *et al.* (2021)提出的 AdaBins 採自適應深度範圍學習機制，將深度範圍劃分為可學習區間(bins)，預測其中心值並線性組合輸出深度圖，以適應不同場景深度變化並提升準確性與穩健性。本研究採用 RGB-D 語義分割模型 DFormer(Yin *et al.*, 2023)結合 AdaBins (Bhat *et al.*, 2021)進行遷移學習，以增強模型在複雜場景下對幾何結構與深度變化的辨識能力。

無人機姿態誤差會影響影像幾何對齊，特別在相機位置與拍攝角度不一致時。儘管二維影像仍為最常用形式(Kluge *et al.*, 2023)，但對於幾何畸變嚴重或結構複雜的場景中，僅依賴二維影像難以修正深度方向誤差，導致傳統拼接方法失效(Liu *et al.*, 2020)。深度資訊輔助二維影像配準可有效消除錯位並保留清晰度(Wang *et al.*, 2024)。Kwon & Lee (2015) 利用深度相機資料推求主法向量並以旋轉變換修正透視變形。本研究則將此幾何校正方法應用於估計深度，以驗證其可行性。

2. 研究方法

2.1 研究流程

本研究以無人機拍攝的橋側 RGB 影像為實測對象，整合多種深度學習模型進行拼接。首先以 AdaBins 進行單張影像深度估計；再透過 DFormer 進行 RGB-D 場景分割，融合色彩與深度資訊區分不同空間區域。為提升拼接的幾何一致性與準確度，本研究設計兩項前處理：依場景分割結果分類影像內容，以及結合深度資訊進行透視投影校正，修正拍攝角度引起的三軸旋轉誤差。最後透過特徵點匹配將不同視角影像拼接，生成具完整性與幾何一致性的橋梁立面全景影像，以利後續狀況觀測與分析。整體研究流程如圖 1 所示。

2.2 橋側數據集建立與相機率定

本研究採用 SONY RX0 數位相機(1 吋 Exmor RS CMOS 感光元件，如圖 2)，具輕巧體積與高解析度影像擷取能力，拍攝解析度為 4800 × 3200 像素。

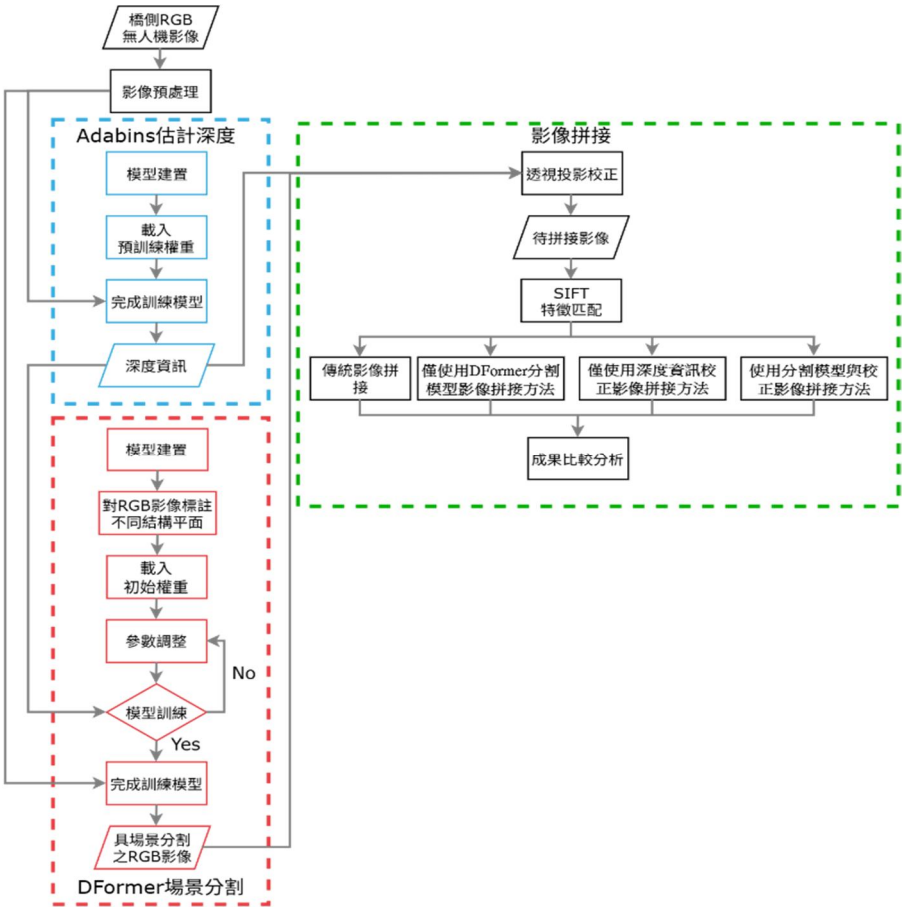


圖 1 研究流程圖

圖 2 Sony RX0 數位相機圖

實驗對象為桃園市大溪區武嶺橋，影像資料為 2021 年 3 月 31 日進行橋梁檢測作業所蒐集，涵蓋南、北兩側立面(如圖 3、4)。共取得 300 張高解析影像作為語義分割模型訓練資料，另選取 7 張側面影像作為測試資料，並應用於後續影像拼接流程研究。

相機率定旨在透過像片上像點的測量，重建拍攝瞬間光束進入鏡頭的精確幾何關係。本研究參考鄒芳諭(2010)提出的非量測型相機率定方法，採用平面棋盤格校準法，從多個視角拍攝黑白棋盤影像(圖 5)，並利用影像中棋盤格交點坐標估算相機的內部參數與畸變係數(表 1)。在取得畸變係數後，即可用來校正影像變形，使畫面更接近真實場景。

圖 3 武嶺橋南側圖

圖 4 武嶺橋北側圖

圖 5 影像率定示意圖

表 1 相機內方位參數表

Sony RX0 內方位參數	
焦距(f_x, f_y)	(3325.9, 3320.81)
像主點位置(X_0, Y_0)	(2344.71, 1583.13)
徑向鏡頭畸變係數(k1,k2)	(-0.0861077, 0.328456)
偏心鏡頭畸變係數(p1,p2)	(0.00156193, -0.00188202)

2.3 AdaBins 預訓練模型與數據集

Bhat *et al.* (2021) 提出 AdaBins 為單張 RGB 影像生成高品質密集深度圖的模型架構。其核心理念在於對傳統編碼器-解碼器(encoder-decoder)輸出的深度分佈進行全域統計分析，並加入學習式後處理模組 AdaBins 於最高解析度下運行，能有效優化深度估計結果。本研究在生成單影像深度資訊時，採用預訓練之 AdaBins 模型，運行環境為 Windows 10 專業版，搭載 12th Gen Intel® Core™ i9-12900F @ 2.40 GHz、NVIDIA GeForce RTX 3080Ti 與 32 GB 記憶體，並於 Python 3.8.0 下執行。AdaBins 原始模型分別於 NYU Depth v2 (Silberman *et al.*, 2012)與 KITTI (Geiger *et al.*, 2012) 資料集完成訓練，可有效由 RGB 影像估計對應深度資訊。本研究選用於 NYU Depth v2 訓練所得之預訓練權重，以生成橋側影像深度資訊。該模型訓練時之設定參數為：初始學習率 1.4×10^{-5}、最大學習率 3.5×10^{-4}、採用 AdamW 優化器、批次大小 16，並於 4 張 NVIDIA® V100(32GB)環境下完成訓練。

2.4 影像標註

RGB-D 場景分割模型的訓練資料需同時包含 RGB 影像、對應語義標註圖與深度影像，且三者必須保持像素對齊，以利模型同時學習色彩特徵、幾何結構與語義資訊，提升分割準確度與場景理解能力。

本研究針對無人機拍攝的 300 張解析度 4800 × 3200 的 RGB 影像進行人工語義標註，共劃分為 16 類語義類別(表 2)。標註作業以 Photoshop 的多邊形套索工具逐一框選並標記語義區域(圖 6)，完成後將 24 位元標註影像轉換為 8 位元單通道格式，並將 16 類標註對應至像素值 0 至 15(圖 7)，以確保一致性與訓練品質。為提升辨識度，將單通道影像像素值乘以 15(圖 8)，最後再將影像縮放至 640 × 480，以符合模型輸入規格。

表 2 各語義類別對應之 RGB 數值表

類別名稱	RGB 數值
橋柱 2	(255,120,20)
橋柱 3	(250,170,30)
管線	(0,60,255)
背景草地	(0,255,50)
鋼板 1	(240,240,110)
鋼板 2	(255,255,0)
鋼板 3	(245,245,75)
背景天空	(0,200,255)

類別名稱	RGB 數值
背景車道	(20,20,20)
橋面板 1	(40,40,40)
橋側邊 1	(60,60,60)
橋側邊 2	(90,90,90)
橋側邊 3	(100,100,100)
橋柱 1	(255,66,0)
隔音板	(200,0,0)
護欄	(140,140,140)

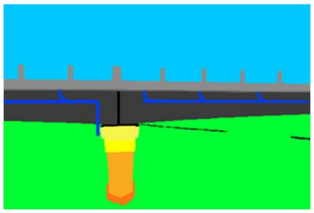


圖 6 原始標註影像

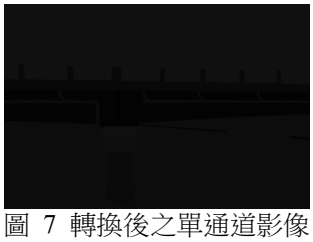


圖 7 轉換後之單通道影像

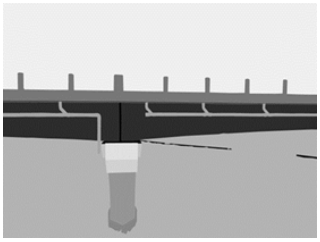


圖 8 像素值 x15 之單通道影像

2.5 DFormer 預訓練模型與數據集

本研究採用 DFormer 預訓練模型(Yin *et al.*, 2023)進行場景分割。透過引入 RGB-D 交互預訓練策略，模型得以克服僅以 RGB 預訓練骨幹在深度圖中易誤編碼幾何資訊的限制，並藉由遷移學習強化橋梁影像之辨識性能。初始訓練數據為 ImageNet-1K(Russakovsky *et al.*, 2015)，並結合 AdaBins 深度估計器(Bhat *et al.*, 2021)生成對應深度資訊，以增強模型的場景理解與泛化能力。

實驗過程使用 300 張橋側 RGB 影像，其中 180 張訓練影像、120 張驗證影像，進行語義標註並產製深度影像，作為遷移學習資料集(圖 9)。模型初始權重引用 DFormer 預訓練模型，並於 Ubuntu 24.04.2 LTS、Intel® Core™ i9-12900F、NVIDIA RTX 3080Ti、32 GB RAM、Python 3.11.10 環境下訓練。參數設置為初始學習率 5×10^{-6}、AdamW 優化器、批次大小 8，並採隨機縮放 {0.5, 0.75, 1, 1.25, 1.5} 完成訓練。

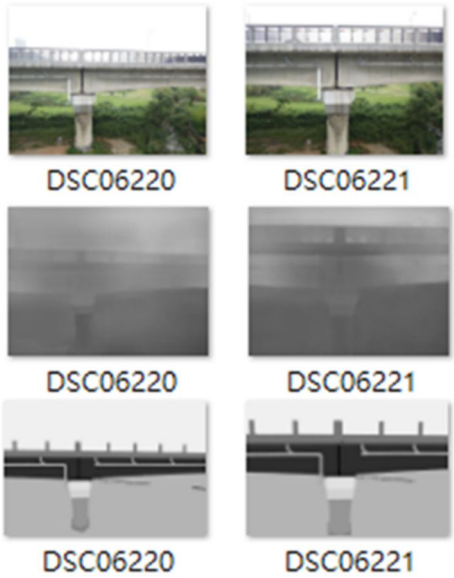


圖 9 用於遷移學習之橋側影像

2.6 DFormer 分割結果後處理

本研究選取 7 張待拼接橋側 RGB 影像作為測試資料，以評估所訓練 RGB-D 場景分割模型之效果。結果顯示，大部分具深度差異的區域能正確辨識，但部分仍存在誤判(圖 10)。為提升拼接中平面區分的準確性，本研究設計後處理方法，主要透過不同窗格尺寸比對測試，調整搜尋窗格與面積閾值修正誤分類，並依鄰近像素標籤重新歸類誤判區域(圖 11)。

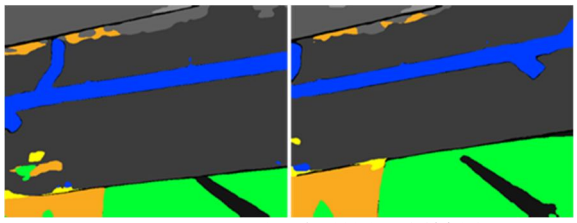


圖 10 DFormer 預測之場景分割結果圖

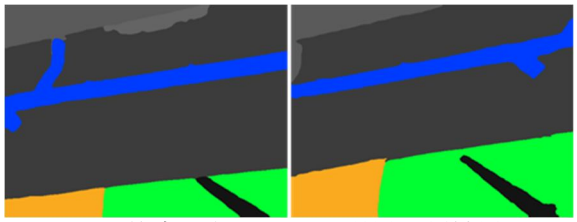


圖 11 後處理之 DFormer 場景分割結果圖

後處理參數包含兩項：

(1) 以每像素為中心擴展為 7×7 視窗，搜尋鄰近不同顏色像素資訊(圖 12)，並排除所有 RGB 值為零的異常區塊。

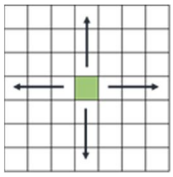


圖 12 誤判區域搜尋視窗示意圖

(2) 測試多組面積閾值 {1000, 2000, 2500, 3000}，逐步調整，直到所有誤判之分類區塊均能正確歸類。

2.7 深度輔助影像校正

傾斜像片為攝影時光軸與鉛垂線形成夾角所獲得的影像。曝光瞬間，即使攝影軸並非刻意傾斜，只要偏離鉛垂線，仍會在像片上產生傾斜現象。為

修正因傾斜造成的幾何變形，需進行幾何校正，使其轉換為同一攝影站位置下的垂直像片(何維信，1995)。像片傾斜主要來自相機在曝光瞬間的姿態變化，其空間取向可由繞 X、Y、Z 軸旋轉的三個角度，即 ω(Roll)、φ(Pitch)、κ(Yaw)所組成的旋轉矩陣 $\boldsymbol{M}$ 表示，用以描述相機在三維空間中的姿態與方向變化，其公式如式(1)。

本研究針對待拼接之橋側 RGB 影像，結合深度資訊進行傾斜校正。由於所選橋側影像屬平面結構，無曲面或大角度起伏，因此僅需將無人機姿態恢復至正攝，使攝影光軸垂直於橋側立面。

(1) 深度影像正規化：影像採用以左上角為原點的 u-v 左手坐標系，深度則以灰階值表示相對距離。故將灰階深度圖正規化至 640×480 影像坐標(圖 13)。

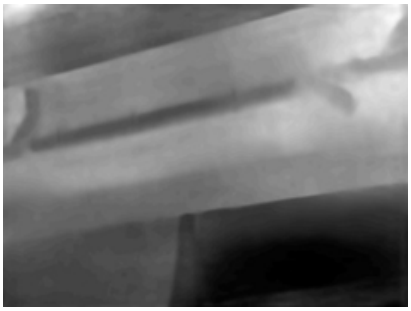


圖 13 深度影像正規化處理圖

(2) 三維點雲重建與語義分類：結合每像素坐標 (u,v) 與灰階深度值，依語義標註篩選主要結構以排除背景干擾(圖 14)。傾斜分析中以深度方向為光軸，分別針對橫軸與縱軸估算傾斜角度(圖 15)，再依據旋轉矩陣將三維點雲校正至無傾斜狀態，完成影像之三維幾何校正(圖 16)。

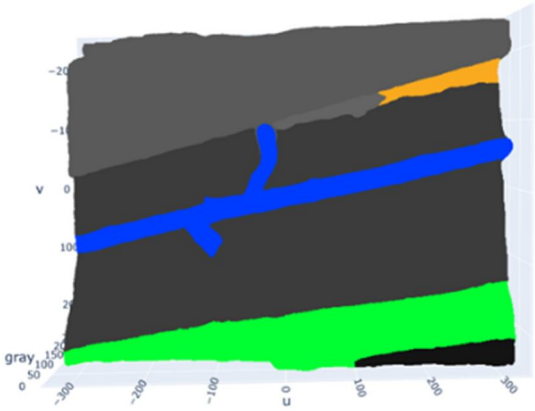


圖 14 語義分割三維點雲俯視圖

(3) 仿射變換與投影映射：利用旋轉矩陣進行反向投影，分別應用至原始 RGB 影像與語義分割影像，並以雙線性內插補齊坐標，避免重疊區域出現稀疏或留白(圖 17)。

$$\boldsymbol{M}=\boldsymbol{M}_{\kappa}\boldsymbol{M}_{\varphi}\boldsymbol{M}_{\omega}=\begin{bmatrix}\cos\kappa & \sin\kappa & 0\\ -\sin\kappa & \cos\kappa & 0\\ 0 & 0 & 1\end{bmatrix}\begin{bmatrix}\cos\varphi & 0 & -\sin\varphi\\ 0 & 1 & 0\\ \sin\varphi & 0 & \cos\varphi\end{bmatrix}\begin{bmatrix}1 & 0 & 0\\ 0 & \cos\omega & \sin\omega\\ 0 & -\sin\omega & \cos\omega\end{bmatrix}=\begin{bmatrix}m_{11} & m_{12} & m_{13}\\ m_{21} & m_{22} & m_{23}\\ m_{31} & m_{32} & m_{33}\end{bmatrix}\quad..(1)$$

式中：

$$m_{11}=cos\,\varphi\,cos\,\kappa$$
$$m_{12}=\cos\omega\sin\kappa+\sin\omega\sin\varphi\cos\kappa$$
$$m_{13}=\sin\omega\sin\kappa-\cos\omega\sin\varphi\cos\kappa$$
$$m_{21}=-\cos\varphi\sin\kappa$$
$$m_{22}=\cos\omega\cos\kappa-\sin\omega\sin\varphi\sin\kappa$$
$$m_{23}=\sin\omega\cos\kappa+\cos\omega\sin\varphi\sin\kappa$$
$$m_{31}=\sin\varphi$$
$$m_{32}=-\sin\omega\cos\varphi$$
$$m_{33}=\cos\omega\cos\varphi$$

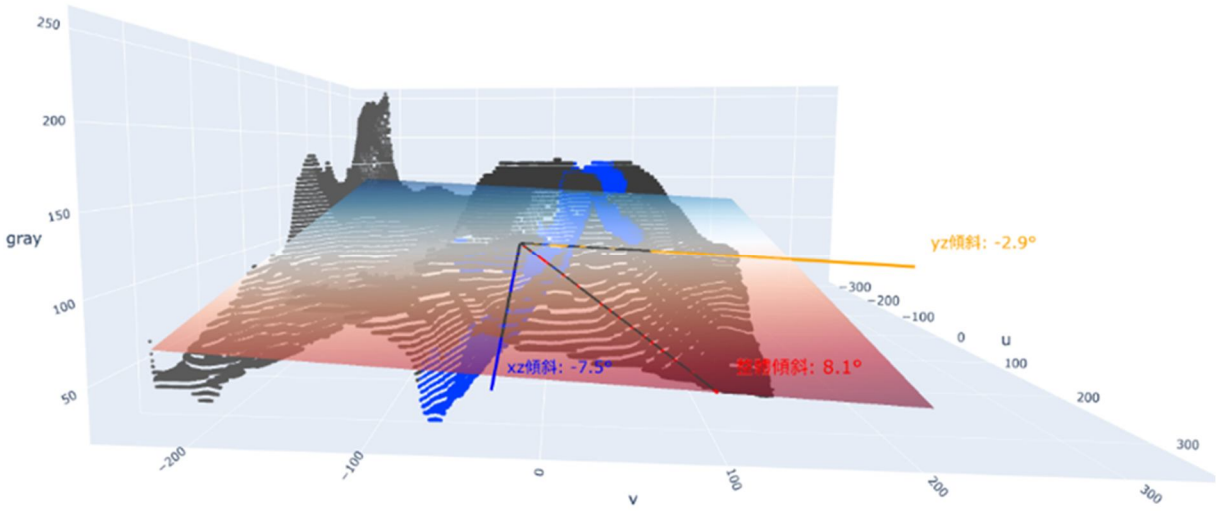


圖 15 影像整體傾斜方向與 ω(Roll)、φ(Pitch)圖

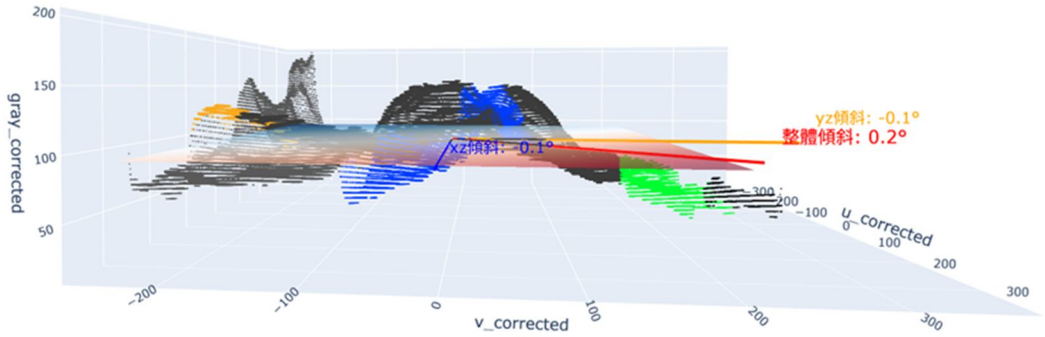
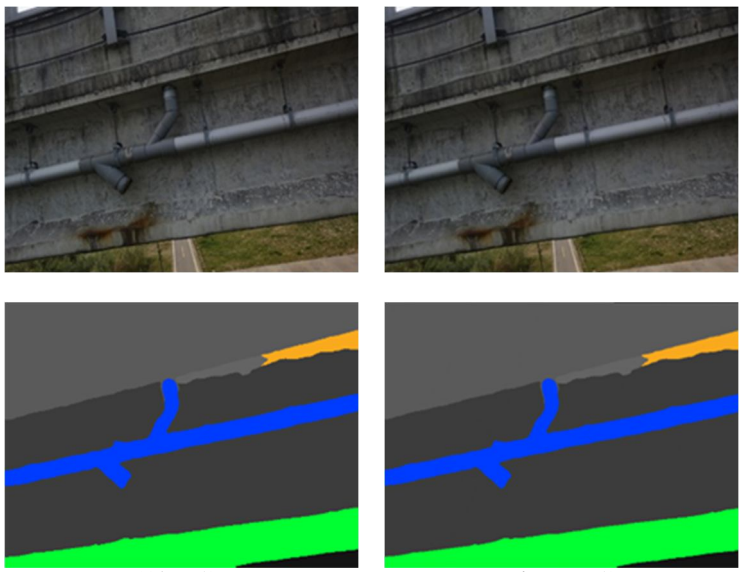


圖 16 校正後影像整體傾斜方向與 ω(Roll)、φ(Pitch)圖



原始影像　　校正影像

圖 17 RGB 與語義分割影像傾斜校正結果圖

2.8 影像拼接

在多張影像的連續拼接過程中，以每組影像的左影像作為參考，固定其坐標系；尺度基準則以第一張影像為依據。相鄰影像之間的重疊率保持在60%以上，以確保具備足夠的有效重疊區域。透過關鍵點偵測與匹配，計算單應性矩陣(Homography Matrix)，將右影像對齊至參考影像，逐步完成拼接。

2.8.1投影變換

又稱為單應性變換(Homography)，指將影像由原始視角轉換至另一視角的操作，可用於產生立體感或修正因拍攝角度造成的透視畸變。在攝影變換的分解中，透視變換描述了影像於投影平面上的幾何對應關係，並可透過一組透視矩陣進行數學建模，其公式如式(2)：

$$\boldsymbol{H}=\begin{bmatrix} scos\theta & -ssin\theta & t_x \\ ssin\theta & scos\theta & t_y \\ 0 & 0 & 1 \end{bmatrix}\begin{bmatrix} 1 & k & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}\begin{bmatrix} \lambda & k0 & 0 \\ 0 & 1/\lambda & 0 \\ 0 & 0 & 1 \end{bmatrix}\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ v1 & v2 & v \end{bmatrix} \quad \text{..... (2)}$$

也可表示為式(3)：

$$\boldsymbol{H}=\boldsymbol{H}_S\boldsymbol{H}_A\boldsymbol{H}_P= \begin{bmatrix} sR & t/v \\ 0^T & 1 \end{bmatrix}\begin{bmatrix} k & 0 \\ 0^T & 1 \end{bmatrix}\begin{bmatrix} 1 & 0 \\ v^T & v \end{bmatrix}=\begin{bmatrix} \boldsymbol{A} & t \\ v^T & v \end{bmatrix} \text{... (3)}$$

2.8.2SIFT 特徵點提取

尺度不變特徵轉換(Scale-Invariant Feature Transform, SIFT)由 Lowe 於 1999 年提出，並於 2004 年完整發表，用於偵測與描述影像中的局部特徵點的演算法，具備尺度與旋轉不變性。其流程包括：首先透過高斯模糊構建尺度空間，並以高斯差分(Difference of Gaussian, DoG)金字塔進行多尺度處理，在其中搜尋局部極值點作為潛在關鍵點。為提升抗噪性與穩定性，進一步以三維二次函數擬合進行位置與尺度細化。隨後根據局部特徵為每個關鍵點分配方向，確保描述子具旋轉不變性。最後，將特徵點鄰域劃分為 4×4 區塊，生成關鍵點描述子，並透過比對參考影像與觀測影像的描述子集合完成特徵匹配。

2.8.3BF 特徵點匹配

完成特徵點擷取後，透過特徵匹配演算法辨識具有高相似性的點對，完成影像間的配對與拼接。本研究於特徵匹配階段中採用 OpenCV 函式庫中的 Brute-Force(BF)匹配器。

BF 匹配器，又稱暴力匹配法，是一種簡單的二維特徵匹配方式，是將一幅影像中的每個特徵點描述子，逐一與另一幅影像中所有描述子比較，並依距離度量選出匹配點。常用策略包含兩種：

(1) 最佳匹配：選取距離最近的單一描述子作為對應點。
(2) *k*-近鄰匹配：返回前 k 個最近的匹配點，由使用者指定 k 值，再透過後處理進行過濾。

暴力匹配在處理大量特徵點時效率較低，但其匹配準確度高、結果穩定，特別適合特徵點較少或對準確性要求高的應用場景。

本研究參考 Kao *et al.* (2024) 方法，在影像連續拼接流程中，先以語義分割區分不同空間平面，再利用 SIFT 偵測特徵點，於原始 RGB 影像計算描述子，並透過 BF 演算法進行匹配。針對每一張目標影像，依據與前一張參考影像的匹配點對估算投影轉換矩陣 (Homography Matrix)，完成幾何對齊與拼接(圖 18)。整體流程包括：

(1) 影像分割：使用語義分割模型劃分不同空間平面。
(2) 特徵點提取：在目標與參考影像的相同平面內進行 SIFT 偵測。
(3) 描述子計算：以原始 RGB 影像生成描述子。
(4) 特徵匹配：利用 BF 匹配器建立兩影像間的特徵點對應。
(5) 變換參數估算：計算投影轉換矩陣。
(6) 影像對齊與拼接：進行幾何變換，使影像逐一對齊並拼接成完整全景圖。

2.9 評估指標

本研究從兩個面向量化評估拼接效果：一為 RGB-D 場景分割模型預測結果與實際標註之重疊程度；另一為拼接後全局影像的一致性與對齊品質。

2.9.1模型性能評估指標

針對 RGB-D 場景分割模型，透過混淆矩陣(表 3)計算精確度(Precision)、召回率(Recall)、F1-score 與交並比(IoU)，量化預測區域與標註資料的重疊一致性。

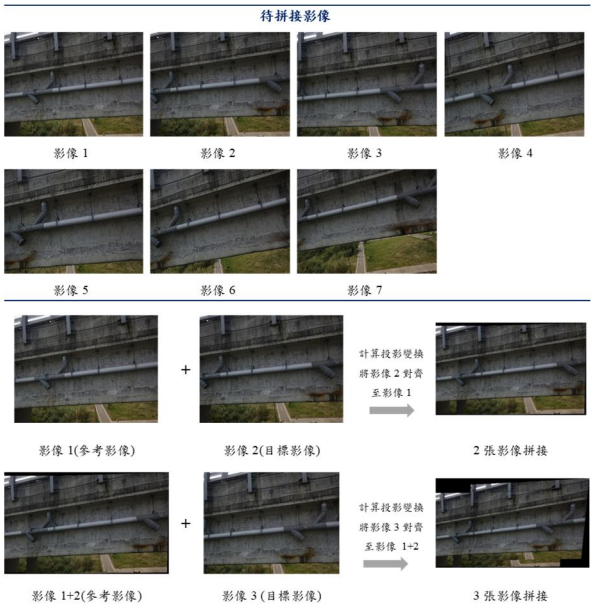


圖 18 連續影像拼接流程示意圖

表 3 分類模型之混淆矩陣評估指標表

	實際為正類	實際為負類
預測為正類	TP	FP(誤報)
預測為負類	FN(漏報)	TN

(1) 精確度(Precision)：衡量模型在預測為正類時的正確比例，如式(4)。

$$Precision = \frac{TP}{TP+FP} \quad (4)$$

(2) 召回率(Recall)：衡量模型對正類樣本的涵蓋能力，如式(5)。

$$Recall = \frac{TP}{TP+FN} \quad (5)$$

(3) F1-score：為精確率(Precision)與召回率(Recall)的調和平均數，適用於樣本不平衡的分類任務中，如式(6)。

$$F1 = \frac{2TP}{2TP+FP+FN} = 2 * \frac{Precision*Recall}{Precision+Recall} \quad (6)$$

(4) 交並比(IoU)： 衡量預測區域與真實標註區域的交集與並集比例，介於 0 至 1，數值越高表示相似度越高，如圖 19 所示。

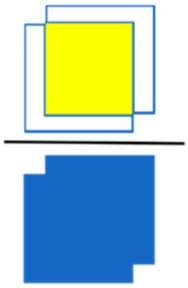


圖 19 交並比(IoU)計算示意圖

2.9.2結構相似性指數

影像拼接品質評估關注錯位情形及亮度與顏色一致性(Zhou *et al.*, 2017)。本研究採用結構相似性指數(SSIM)評估全局影像的對齊與內容一致性。相較於均方誤差(MSE)與峰值訊號雜訊比(PSNR)，SSIM 以人類視覺感知為基礎，綜合考量亮度、對比與結構，能更準確反映影像間具感知差異的誤差，具有較高的主觀一致性與實用價值(Sara *et al.*, 2019)。SSIM 值介於 －1 至 1，越接近 1 表示兩張影像越相似。SSIM($\boldsymbol{x},\boldsymbol{y}$) 為評估兩張影像相似程度的指標，如式(7)~(10)所示。

$$SSIM(x,y) = [\boldsymbol{L}(x,y)^{\alpha}][\boldsymbol{C}(x,y)^{\beta}][\boldsymbol{S}(x,y)^{\gamma}] \quad (7)$$

$$\boldsymbol{L}(x,y) = \frac{2\mu_x\mu_y+\boldsymbol{c1}}{\mu_x{}^2+\mu_y{}^2+\boldsymbol{c1}} \quad (8)$$

$$\mathbf{C}(x,y) = \frac{2\sigma_x\sigma_y+\boldsymbol{c2}}{\sigma_x{}^2+\sigma_y{}^2+\boldsymbol{c2}} \quad (9)$$

$$\mathbf{S}(x,y) = \frac{2\mu_{xy}+\boldsymbol{c3}}{\sigma_x\sigma_y+\boldsymbol{c3}} \quad (10)$$

式中：

$\boldsymbol{L}(\boldsymbol{x},\boldsymbol{y})$表示亮度項

$\mathbf{C}(\boldsymbol{x},\boldsymbol{y})$表示對比度項

$\mathbf{S}(\boldsymbol{x},\boldsymbol{y})$表示結構項

α、β、γ 分別為各項的加權係數

μ_x與μ_y為兩張影像的像素平均值

$\sigma_x{}^2$與$\sigma_y{}^2$為兩張影像的方差

$\sigma_x\sigma_y$為兩張影像之間的協方差

$\boldsymbol{c1}$、$\boldsymbol{c2}$、$\boldsymbol{c3}$ 用於避免分母為零

3. 研究成果與分析

3.1 DFormer 訓練成果與分析

本章節將分析模型訓練結果，以 300 張影像作為訓練數據，隨機劃分為訓練集 180 張與驗證集 120 張，並依檔名對應 RGB、深度 (Depth) 與語義標註 (Label) 影像，以利資料載入與訓練。模型訓練前載入預訓練權重作為初始參數，並於單張 NVIDIA RTX 3080 GPU 環境下進行。訓練影像統一調整為 640 × 480 像素，初始學習率設為 6e-5，隨訓練逐步衰減；優化器採 AdamW，動量 0.9，權重衰減 0.01。為提升泛化能力，訓練中採多尺度[0.5, 0.75, 1, 1.25, 1.5]與隨機翻轉策略，每批次含 8 張影像，共訓練 500 個 epoch。

考量 GPU 資源限制，本研究選用輕量級 DFormer-T (Tiny) 作為編碼器，以兼顧效率與可行性。圖 20 顯示訓練過程的損失函數變化。

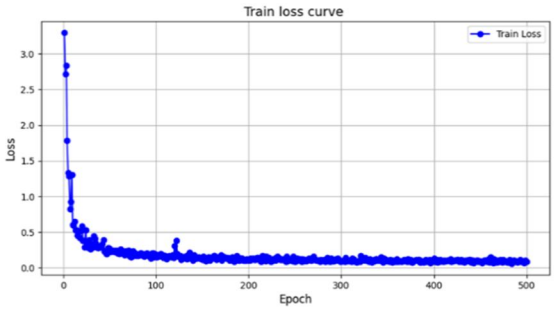


圖 20 訓練階段之損失函數曲線圖

在驗證階段，本研究以平均交並比 (mIoU) 評估模型效能。如圖 21 所示，隨著訓練過程中損失函數逐漸收斂，mIoU 指數亦呈現穩定上升趨勢，最終驗證結果達到 86.44%，顯示模型具良好的語義分割準確性與收斂表現。

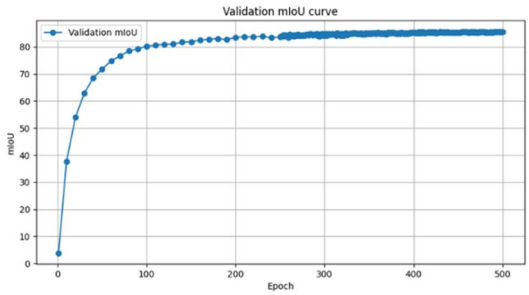


圖 21 驗證階段之 mIoU 曲線圖

並針對完成訓練之模型進行多項性能指標評估，包含各語義分割類別的交並比(IoU)(圖 22)、精確度(Precision)與召回率(Recall)(圖 23)以及 F1-score，如表 4 所示。

圖 22 各類別 IoU 分數曲線圖

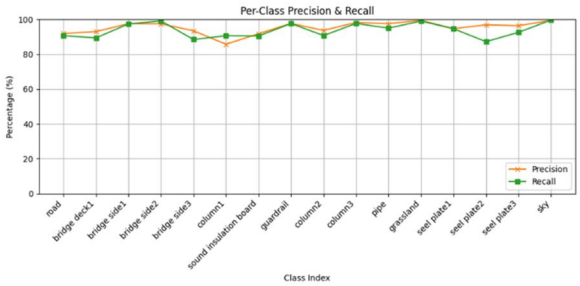


圖 23 各類別精度與召回率分數曲線圖

表 4 模型訓練各語義類別之平均精度指標

mIoU	mAcc	Mean Precision	Mean Recall	mF1
86.44%	92.1%	91.24%	92.11%	91.56%

3.2 影像拼接成果比較與分析

3.2.1相鄰影像拼接成果

進行連續拼接前，先對 DFormer 分割影像與深度校正影像進行兩張影像拼接測試，並與 Zaragoza *et al.* (2013) 所提出的 APAP 演算法比較。實驗結果顯示，APAP 在低紋理區域因關鍵點不足而受限(圖 24)，使扭曲模型估算不準確，導致影像錯位(圖 25)；相較之下，結合 DFormer 分割搭配 Kao *et al.* (2024) 所提方法，利用原始 RGB 影像提取描述子，可有效提升對齊精度與拼接品質(圖 26)。進一步將語義分割資訊導入拼接流程(圖 27)，能改善大視差下的特徵對應，提升穩定性與一致性；若先進行傾斜校正再結合分割結果(圖 28)，則能進一步減少幾何變形，確保拼接的準確性與完整性。

圖 24 APAP 演算法相鄰 RGB 影像特徵點匹配展示圖

圖 25 APAP 演算法之相鄰 RGB 影像拼接展示圖

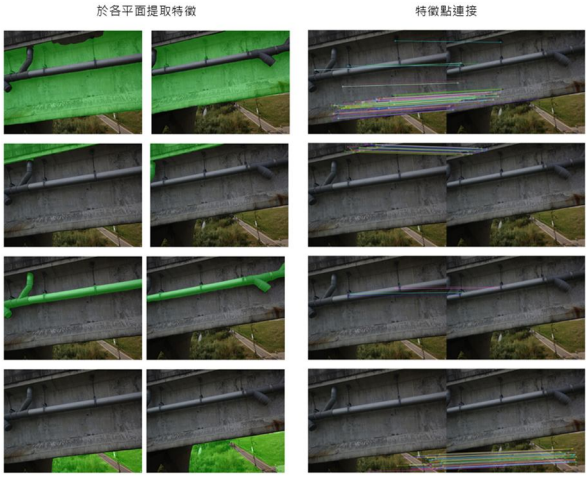


圖 26 僅使用 DFormer 分割影像特徵點匹配展示圖

圖 27 僅使用 DFormer 模型分割之相鄰 RGB 影像拼接展示圖

圖 28 使用分割模型與校正影像方法之相鄰 RGB 影像拼接展示圖

3.2.2相鄰影像拼接精度評估

針對三種不同預處理方法的相鄰拼接結果進行結構相似性指數(SSIM)評估，其分析結果見表 5、圖 29。

表 5 相鄰影像拼接之評估指標比較表

影像對	傳統方法	僅使用 DFormer 分割模型方法	僅使用深度資訊校正影像方法	使用分割模型與校正影像方法
影像 1& 影像 2	0.4896	0.5719	0.5204	0.6136
影像 2& 影像 3	0.4943	0.6228	0.5305	0.6681
影像 3& 影像 4	0.4012	0.6402	0.4312	0.6850
影像 4& 影像 5	0.4768	0.6292	0.5069	0.6924
影像 5& 影像 6	0.6200	0.6664	0.6730	0.7306
影像 6& 影像 7	0.5669	0.6323	0.5977	0.6947
Average	0.5081	0.6271	0.5433	0.6807

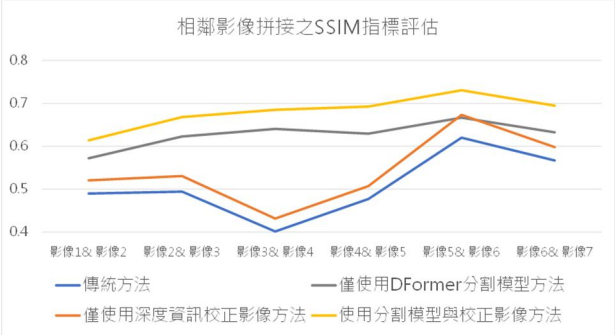


圖 29 相鄰影像拼接之評估指標比較折線圖

3.2.3連續影像拼接成果

本研究以 7 張橋側影像比較三種拼接策略與傳統方法之差異，分別為：僅用 DFormer 分割模型、僅用深度校正影像，以及結合分割與校正影像，以評估橋梁影像拼接的準確性。

(1) 傳統方法：透過 SIFT 偵測特徵點與描述子，使用 BF 匹配並結合 RANSAC 排除異常後估算單應矩陣完成拼接。但因僅依賴 RGB 資訊，當影像存在深度差異時，易產生對齊誤差，影響拼接品質(圖 30)。

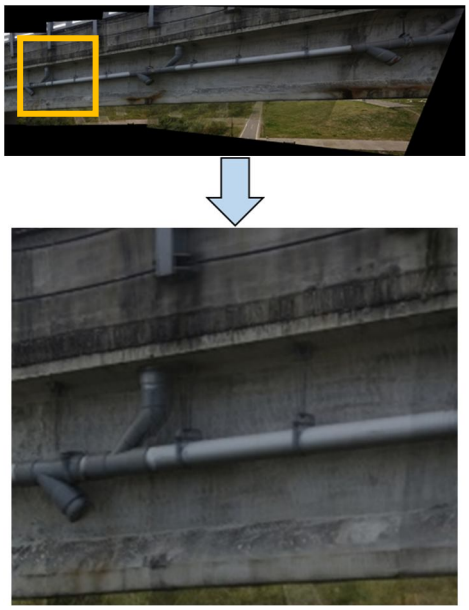


圖 30 傳統方法之連續影像拼接成果圖

(2) 僅用 DFormer 分割模型：透過語義分割辨識橋梁主體與背景，於橋梁結構上可達較高對齊精度。但因拍攝角度與深度差異，背景區域易產生視差與局部錯位。藉由分割遮罩排除不具幾何一致性的背景，可避免誤匹配並提升拼接準確度(圖 31)。

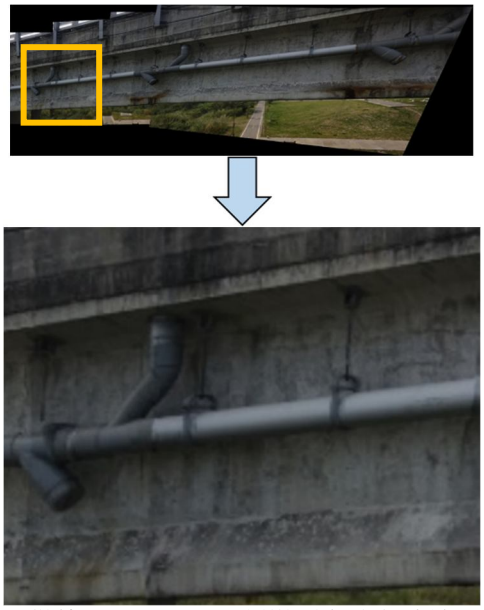


圖 31 僅使用 DFormer 分割模型方法之連續影像拼接成果圖

(3) 僅用深度校正影像：透過深度資訊進行幾何校正，可補償無人機姿態誤差，降低拼接中的幾何變形，並提升連續拼接的對齊穩定性(圖 32)。

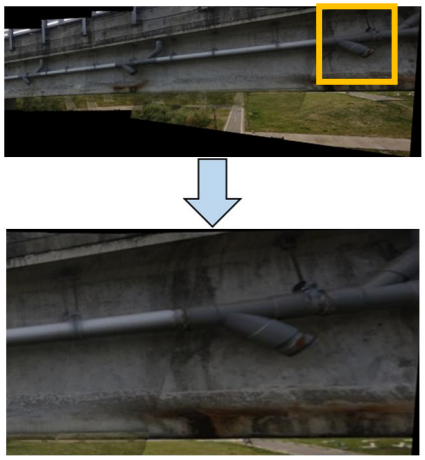


圖 32 僅使用深度資訊校正影像方法之連續影像拼接成果圖

(4) 結合分割模型與校正影像：先利用深度資訊估算拍攝姿態並計算旋轉矩陣，再對 RGB 與 RGB-D 影像進行傾斜校正，修正因角度與視差造成的幾何扭曲。此方法不僅提升跨視角影像的對齊準確度，也保留分割資訊對橋梁平面的辨識能力，進而改善拼接的一致性與精度(圖 33)。

圖 33 使用分割模型與校正影像方法之連續影像拼接成果圖

3.2.4連續影像拼接成果分析

本實驗採用以序列首張影像作為尺度基準，後續影像依序進行投影變換與對齊。由於未進行全局優化，誤差會隨拼接順序逐步累積，導致序列後段出現明顯的幾何變形與對位偏差，特別在長距離拼接中更為顯著，顯示缺乏尺度統一與誤差抑制機制時，拼接結果易受累積誤差影響。

4. 結論

4.1 結論

本研究提出之影像拼接方法無需繁瑣的三維建模，以橋側影像驗證其效能，有效降低建模時間

與硬體需求。本研究方法結合單影像深度估計與 RGB-D 場景分割模型，並利用深度資訊進行傾斜校正，修正拍攝姿態造成的幾何變形，提升複雜場景下的拼接品質與對齊準確性。綜合實驗結果，結論如下四點：

(1) 透過遷移學習提升 RGB-D 場景分割模型於橋梁場景的辨識能力，訓練過程損失函數穩定收斂，驗證階段 mIoU 提升至 86.44%，最終 mAcc 達 92.1%。結合 AdaBins 生成之深度圖，有效區分不同空間深度物體並提升語義辨識準確度，另以後處理機制修正誤判區域，進一步改善分割效果。

(2) 透過深度資訊輔助 RGB 影像進行幾何校正，修正因相機傾斜造成的影像扭曲，有效降低拼接時的幾何變形，提升對齊準確性與拼接品質。

(3) 傳統以 SIFT 提取特徵點結合 APAP 網格扭曲方法對參數較為敏感，易導致拼接錯位；本研究透過語義分割與深度資訊輔助的影像校正策略，有效提升複雜場景下的拼接精度與穩定性。以結構相似性指數(SSIM)量化評估拼接品質，實驗結果顯示，三種預處理方法之平均 SSIM 值分別為 0.6271、0.5433、0.6807，皆顯著高於傳統方法之 0.5081，顯示本方法在視覺一致性上具明顯優勢。

(4) 本研究產製之橋側立面全景影像可作為橋梁檢測與管理的記錄底圖，特別適用於長裂縫等無法以單張影像完整呈現的作業需求，有助提升結構診斷效率與作業系統化程度。

4.2 建議

(1) 本研究針對語義分割中的誤判區域，採用逐步閾值調整結合鄰近像素分類進行修正，雖有效提升分割品質，但在大範圍或高複雜度場景中處理成本較高。建議未來導入自動化閾值選取或機器學習方法，以提升效率與泛化能力。

(2) 本研究直接使用預訓練之 AdaBins 模型生成深度資訊，建議未來加入混凝土橋梁影像進行遷移學習，以強化其在結構場景下的深度估計準確性，進一步提升 RGB-D 語義分割精度。

(3) 本研究所選用之待拼接影像以連續影像中的第一張作為尺度參考，惟影像間存在尺度不一致的情況，可能導致誤差累積。建議未來可透過轉換公式統一尺度，或採用等距離拍攝方式以維持影像尺度一致性，進一步提升拼接穩定性與準確性。

(4) 本研究驗證場域為混凝土橋梁側面影像，建議未來擴展至更複雜環境與多目標語義分類任務，拓展於智慧建築監測與自動巡檢等應用場景。

參考文獻

何維信，1995。航空攝影測量學，大中國圖書公司。[Ho, W.H., 1995. Photogrammetry, The Great China Book Co., Ltd., Taiwan, ROC. (in Chinese)]

鄒芳諭，2010。以非量測性相機進行近景攝影測量，國立交通大學土木工程學系碩士論文。[Tsou, F.Y., 2010. Analysis of close-range photogrammetry by using non-metric camera, Master Thesis, National Chiao Tung University, Taiwan, ROC. (in Chinese)]

Bhat, S.F., Alhashim, I., and Wonka, P., 2021. Adabins: Depth estimation using adaptive bins, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, pp. 4009-4018, DOI: 10.1109/CVPR46437.2021.00400.

Cai, W., Du, S., and Yang, W., 2023. UAV image stitching by estimating orthograph with RGB cameras, Journal of Visual Communication and Image Representation, 94: 103835, DOI: 10.1016/j.jvcir.2023.103835.

Geiger, A., Lenz, P., and Urtasun, R., 2012. Are we ready for autonomous driving? the KITTI vision benchmark suite, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, pp. 3354-3361, IEEE, DOI: 10.1109/CVPR.2012.6248074.

Kao, S.P., Lin, J.S., Wang, F.L., and Hung, P.S., 2024. A large-crack image-stitching method with cracks as the regions of interest, Infrastructures, 9(4): 74, DOI: 10.3390/infrastructures9040074.

Kekec, T., Yildirim, A., and Unel, M., 2014. A new approach to real-time mosaicing of aerial images, Robotics and Autonomous Systems, 62(12): 1755-1767, DOI: 10.1016/j.robot.2014.07.010.

Khan, F., Salahuddin, S., and Javidnia, H., 2020. Deep learning-based monocular depth estimation methods—A state-of-the-art review, Sensors, 20(8): 2272, DOI: 10.3390/s20082272.

Kluge, M., Weyrich, T., and Kolb, A., 2023. Progressive refinement imaging with depth-assisted disparity correction, Computers & Graphics, 115: 446-460, DOI: 10.1016/j.cag.2023.07.036.

Kwon, S.K., and Lee, D.S., 2015. Correction of perspective distortion image using depth information, Journal of Korea Multimedia Society, 18(2): 106-112, DOI: 10.9717/kmms.2015.18.2.106.

Li, J., Wu, D., Jiang, P., Li, Z., and Song, S., 2021. Locally aligned image stitching based on multi-feature and super-pixel segmentation with plane protection, IEEE Access, 9: 168315-168328, DOI: 10.1109/ACCESS.2021.3134887 .

Lin, C.C., Pankanti, S.U., Ramamurthy, K.N., and Aravkin, A.Y., 2015. Adaptive as-natural-as-possible image stitching, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, pp. 1155-1163, DOI: 10.1109/CVPR.2015.7298719.

Liao, K., Liu, G., and Hui, Y., 2013. An improvement to the SIFT descriptor for image representation and matching, Pattern Recognition Letters, 34(11): 1211-1220, DOI: 10.1016/j.patrec.2013.03.021.

Liu, Y.F., Nie, X., Fan, J.S., and Liu, X.G., 2020. Image-based crack assessment of bridge piers using unmanned aerial vehicles and three-dimensional scene reconstruction, Computer-Aided Civil and Infrastructure Engineering, 35(5): 511-529, DOI: 10.1111/mice.12501.

Lowe, D.G., 2004. Distinctive image features from scale-invariant keypoints, International Journal of Computer Vision, 60: 91-110, DOI: 10.1023/B:VISI.0000029664.99615.94.

Ma, W., Wen, Z., Wu, Y., Jiao, L., Gong, M., Zheng, Y., and Liu, L., 2016. Remote sensing image registration with modified SIFT and enhanced feature matching, IEEE Geoscience and Remote Sensing Letters, 14(1): 3-7, DOI: 10.1109/LGRS.2016.2600858.

Rajapaksha, U., Sohel, F., Laga, H., Diepeveen, D., and Bennamoun, M., 2024. Deep learning-based depth estimation methods from monocular image and videos: A comprehensive survey, ACM Computing Surveys, 56(12): 315, DOI: 10.1145/3677327.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., and Fei-Fei, L., 2015. Imagenet large scale visual recognition challenge, International Journal of Computer Vision, 115: 211-252, DOI: 10.1007/s11263-015-0816-y.

Sara, U., Akter, M., and Uddin, M.S., 2019. Image quality assessment through FSIM, SSIM, MSE and PSNR—A comparative study, Journal of Computer and Communications, 7: 8-18, DOI: 10.4236/jcc.2019.73002.

Silberman, N., Hoiem, D., Kohli, P., and Fergus, R., 2012. Indoor segmentation and support inference from rgbd images, in Proceedings of the Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, pp. 746-760, DOI: 10.1007/978-3-642-33715-4_54.

Tang, Z., Zhang, Z., Chen, W., and Yang, W., 2023. An SIFT-based fast image alignment algorithm for

high-resolution image, IEEE Access, 11: 42012-42041, DOI: 10.1109/ACCESS.2023.3270911.

Tu, J., Sui, H., Feng, W., Sun, K., Xu, C., and Han, Q., 2017. Detecting building facade damage from oblique aerial images using local symmetry feature and the Gini index, Remote Sensing Letters, 8(7): 676-685, DOI: 10.1080/2150704X.2017.1312027.

Wang, C., Wang, C., Li, W., and Wang, H., 2021. A brief survey on RGB-D semantic segmentation using deep learning, Displays, 70: 102080, DOI: 10.1016/j.displa.2021.102080.

Wang, T., Huang, H., Cai, Z., Song, J., and Yang, J., 2024. 360° panorama stitching method with depth information: Enhancing image quality and stitching accuracy, The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 48: 191-197, DOI: 10.5194/isprs-archives-XLVIII-4-W10-2024-191-2024.

Xiang, T.Z., Xia, G.S., Bai, X., and Zhang, L., 2018. Image stitching by line-guided local warping with global similarity constraint, Pattern Recognition, 83: 481-497, DOI: 10.1016/j.patcog.2018.06.013.

Yin, B., Zhang, X., Li, Z., Liu, L., Cheng, M.M., and Hou, Q., 2023. Dformer: Rethinking rgbd representation learning for semantic segmentation, arXiv preprint, arXiv:2309.09668, DOI: 10.48550/arXiv.2309.09668.

Zaragoza, J., Chin, T.J., Brown, M.S., and Suter, D., 2013. As-projective-as-possible image stitching with moving DLT, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR), Portland, OR, USA, pp. 2339-2346, DOI: 10.1109/CVPR.2013.303.

Zhou, X., Zhang, H., and Wang, Y., 2017. A multi-image stitching method and quality evaluation, in Proceedings of the 4th International Conference on Information Science and Control Engineering (ICISCE), Changsha, China, pp. 46-50, DOI: 10.1109/ICISCE.2017.20.

Research on Depth-Enhanced Image Stitching for Concrete Bridge Structures

Yu-Ting Lin [1*] Szu-Pyng Kao [2] Feng-Liang Wang [3] Jhih-Sian Lin [4]

Abstract

Image stitching can expand the field of view and eliminate blind spots, but differences in scene depth often lead to parallax and ghosting. To address this, this study integrates single-image depth estimation and semantic segmentation models to establish a façade image stitching workflow for bridges, reconstructing a complete structural appearance map for damage analysis and management. Through transfer learning, a bridge-side image dataset was constructed, and a pretrained RGB-D semantic segmentation model was fine-tuned. The model achieved an mIoU of 86.44%, mAcc of 91.24%, recall of 92.11%, and F1-score of 91.56%, demonstrating stability and generalization capability, while indirectly validating the accuracy of the depth estimation model. To correct geometric misalignment caused by image tilt, depth maps were used to reconstruct point clouds for correction. A stitching accuracy comparison shows that the average SSIM of images corrected with the segmentation model (0.6807) was higher than that of the traditional method (0.5081), confirming the advantages of the proposed approach in terms of accuracy and visual consistency.

Keywords: Bridge Façade Inspection, Image Stitching, Depth Estimation, Semantic Segmentation, Image Geometric Correction

[1] Master, Department of Civil Engineering, National Chung Hsing University
[2] Professor, Department of Civil Engineering, National Chung Hsing University
[3] Assistant Professor, Department of Applied Geoinformatics, Chien Hsin University of Science and Technology
[4] Ph.D., Department of Civil Engineering, National Chung Hsing University
[*] Corresponding Author, Tel: 886-919907056, E-mail: venusborn2023@gmail.com

Received Date: Sep. 03, 2025
Revised Date: Nov. 21, 2025
Accepted Date: Feb. 11, 2026