航測及遙測學刊　第二十九卷　第 4 期　第 211- 230 頁　民國 113 年 12 月
Journal of Photogrammetry and Remote Sensing
Volume 29, No.4, 2024, pp. 211 – 230
DOI：10.6574/JPRS.202412_29(4).0002

深度學習影像特徵匹配應用於無人機影像視覺定位

鄒來翰 [1*]　林昭宏 [2]

摘要

當無人機 (unmanned aerial vehicle, UAV) 配備之定位及定向設備無作用時，可使用影像視覺定位技術單以影像共軛點進行空間後方交會推導載具外方位。本研究提出影像視覺定位流程，並改善使用深度學習模型特徵點匹配時因影像間平面旋轉而匹配成功率大幅降低地問題。加入資料擴增隨機旋轉影像，以特徵萃取模型提取特徵後輸入匹配模型學習。另外，透過提出內插法以及可學習參數法將原本用於匹配之特徵描述符替換為傳統特徵描述符，使其具有旋轉不變性。萃取影像中之特徵點並進行匹配後，可用一般傳統攝影測量空間後方交會求解位於載具上的相機 6 個外方位元素，進行載具定位。經本文影像視覺定位流程，解算外方位平面位置誤差最佳可達 3 m、姿態角誤差最佳可達 1.3°。

關鍵詞：深度學習、特徵萃取、影像匹配、視覺定位、旋轉不變性

1. 前言

現代大部分無人機都配備了全球定位系統 (Global Positioning System, GPS) 和慣性測量單元 (Inertial measurement unit, IMU) 等定位及定向設備，以確保飛行過程中的精確導航和定位。然而，在戰爭時期或存在遮蔽物的環境中，這些定位系統可能會失效，因此需要使用影像視覺地形輔助定位技術來解決問題。該技術是一種不依賴於傳統定位系統的技術，僅通過搭載的攝影鏡頭對地面拍攝取得即時航拍影像，並與參考影像匹配特徵共軛點，進行空間後方交會來推導載具的外方位元素。這種技術在無人機航拍影像應用中尤為重要，特別是在 GPS 信號受阻或干擾的情況下，能夠提供可靠的定位和導航解決方案。無人機的自主導航依賴於精確的路徑規劃演算法，目前許多基於圖論 (graph) 的演算法被廣泛應用於此領域，例如 Dijkstra (Luo *et al.*, 2020) 以及 A* (Ju *et al.*, 2020) 等演算法，都被廣泛應用於搜尋最短路徑。隨後，D* Lite 演算法 (Chang *et al.*, 2023) 作為 Dijkstra 和 A*的改進版本問世，進一步提升了路徑規劃的效率和靈活性。影像視覺定位技術使無人機能夠在任務執行過程中即時計算自身位置，結合路徑規劃技術可實現自主飛行準確抵達目的地。

有關無人機影像補助定位，根據研究 (黃敬群及黃偉立，2012、Lin & Medioni, 2007) 所提出的方法，透過影像匹配計算影像之間的對應關係，便可將一序列的無人機影像對應至衛星影像上，之後可透過座標系統的轉換進行無人機位置的計算。Chen *et al.* (2021a) 提出在無 GPS 的環境中快速且穩定的對無人機進行地理定位之框架，該研究將衛星影像事先透過深度學習模型編碼為全域描述符並封裝成資料庫裝載於無人機上，無人機拍攝影像後也會透過相同的方式進行編碼，並透過描述符於資料庫中搜尋相似之影像，並使用深度學習影像匹配演算法對無人機影像與搜索到的衛星影像進行匹配後求解無人機外方位，該方法將能夠運算的先執行以簡化運算的負荷，以達到近乎實時的定位方式。以上方式皆有使用到影像匹配技術，也說明使用無人機進行影像視覺定位時，與具地理座標之衛星影像之間的轉換關係，需透過影像匹配技術達成，如何獲取足夠且穩固的特徵點為研究的重要目標。

[1] 國立成功大學測量及空間資訊學系 碩士
[2] 國立成功大學測量及空間資訊學系 教授
[*] 通訊作者, E-mail: ha095863@gmail.com

收到日期：民國 113 年 09 月 16 日
修改日期：民國 113 年 10 月 08 日
接受日期：民國 113 年 10 月 21 日

無人機的定位由六個外方位元素 (包括位置和姿態) 決定，可由空間後方交會解算獲得。空間後方交會解算需要特徵共軛點，主要依賴於影像特徵萃取和匹配技術。傳統的特徵偵測和描述演算法如 SIFT (Low, 2004)、SURF (Bay *et al.*, 2006) 等透過計算灰度值梯度萃取影像中特徵點的位置及描述特徵結構。特徵匹配負責比較前述萃取之特徵點描述符並找出對應的匹配關係，如 Brute-Force Matching，直接計算所有特徵點描述符之間的距離，並選擇距離最小的那一對作為匹配。FLANN (Muja & Lowe, 2009) 則專為高維度數據的最近鄰搜索。然而，傳統方法由於使用影像灰度值，容易受光影、亮度等影響，使得它們在面對複雜場景時性能和精度仍然存在一些局限。隨著時間的推移和技術的進步，深度學習相關的研究得到了迅速發展，並在計算機視覺領域取得了顯著的成果。深度學習 (Deep Learning) 方法能夠自動從數據中學習特徵表示，提高匹配的精度和穩定性，且因其可學習參數的特性，能夠針對特定領域進行特化和改善。如針對無人機航拍影像的應用，可用航拍影像訓練資料集使模型對於航拍影像的推演任務表現提升，使得深度學習方法於特徵萃取與匹配之表現相比於傳統方法更加穩定且精度較高 (Ma *et al.*, 2021)。眾多特徵提取之深度學習模型，其中監督式的方法，如 Lift (Yi *et al.*, 2016)、Tilde (Verdie *et al.*, 2015) 等，可能使得模型受限於人為錨點的設計而難以提出新的特徵點。自監督式學習的特徵點提取及描述模型 SuperPoint (DeTone *et al.*, 2018)，則只需要簡單的人為幫助進行預訓練。特徵匹配之深度學習模型，NCNet (Rocco *et al.*, 2020a)、Sparse-NCNet (Rocco *et al.*, 2020b) 等透過相關性比對匹配特徵。SGMNet 方法 (Chen *et al.*, 2021b) 引入種子點來引導圖匹配的過程，種子點是預先已知的匹配點，這些點在匹配過程中可以作為引導，幫助模型學習圖中其他點的對應關係。LoFTR 方法 (Sun *et al.*, 2021) 是一種不需描述符之匹配法，不需兩階段先提取特徵再匹配，模型可直接萃取全局資訊並產生密集匹配。SuperGlue (Sarlin *et al.*, 2020) 則利用了自注意力與交叉注意力機制來分析空間分布關係增強特徵點之間的匹配。LightGlue (Lindenberger *et al.*, 2023) 為 SuperGlue 改進版，藉由對匹配難度自適應匹配提出更快速的匹配效率。由於 SuperGlue 模型相較其他模型輕量適合置入無人機，且經實測後 LightGlue 應用於航拍影像效率與 SuperGlue 相差甚小，因此本研究最終決定使用 SuperPoint 搭配 SuperGlue 的匹配流程，該改善模型訓練流程爾後也可應用於其他模型。

無人機執行任務時，以影像視覺定位方法決定自身位置，由於飛行方向的變化，不同航帶之間常會出現影像之間平面旋轉問題，導致使用深度學習匹配過程中的困難，因為影像中的特徵點在旋轉後的位置和方向都會發生變化。雖已有傳統匹配方法具有對旋轉的抵抗能力 (如 SIFT)，然而解算得外方位通常誤差較大且不穩定。欲使用深度學習匹配簡單之做法為每次旋轉90°無人機拍攝影像匹配一次，四個方向上匹配點數最多即作為匹配成果，雖可解決問題，但會大幅增加計算資源和時間的消耗。若能使用深度學習方法同時克服旋轉問題導致的匹配困難，可在獲得更精確定位之同時大幅降低所需計算時間與資源，使得解算更有效率。

本研究提出影像視覺地形輔助定位技術之自動化流程，將深度學習應用於特徵點匹配，並改良原本影像特徵萃取和匹配時需先對影像旋轉的問題，使得影像特徵匹配流程只需進行一次，減少計算資源和時間的消耗，使得影像視覺地形輔助定位技術更有效率。

2. 研究方法

本研究探討無人機地形輔助視覺定位應用如圖 1 所示，其流程可大致分為三個步驟：(1) 資料前處理，找出無人機影像對應參考影像位置與解析度一致化；(2) 匹配參考影像與無人機影像共軛特徵點；(3) 搭配 DSM 列出共線式，透過空間後方交會解算外方位。

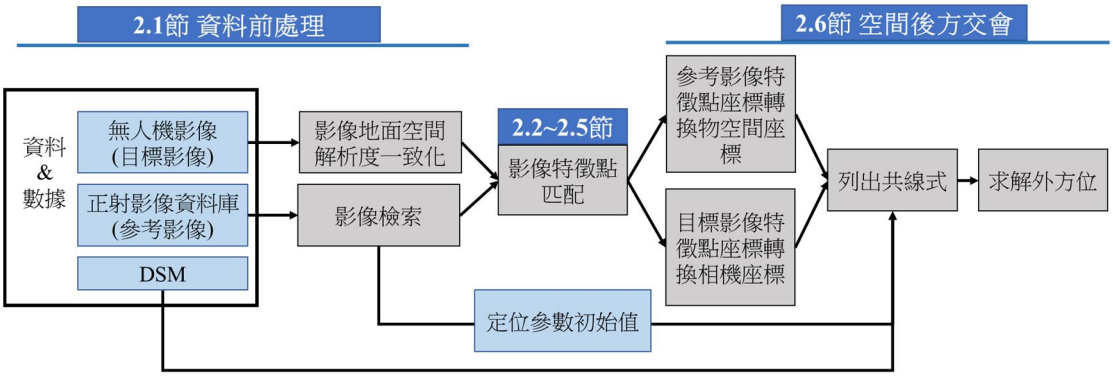


圖 1 無人機地形輔助視覺定位詳細流程

2.1 資料前處理

由於無人機影像與參考影像之間像幅與拍攝範圍差距大，故匹配前須將參考影像裁切至適當大小且涵蓋地面位置與無人機影像重疊。此處可以影像檢索方法實現，如圖 2 所示。先將參考影像裁切成多張大小與無人機影像相當，再使用輕量化的模型如 MobileNet (Sinha & El-Sharkawy, 2019)，將裁切之參考影像萃取特徵圖並接上 NetVLAD (Arandjelovic *et al.*, 2016) 對影像產生一個全域特徵描述符 (Global Descriptor)，該描述符對整張影像的特徵進行描述，而後將所有參考影像之全域描述符組成一資料庫存入無人機機載記憶體中。無人機任務進行時可透過產生之無人機影像全域描述符於資料庫中檢索相似之參考影像，並選出前 N 個相似之參考影像，以利後續影像特徵點匹配步驟；同時也可根據檢索結果提供無人機影像之定位參數初始值，以利後續空間後方交會。

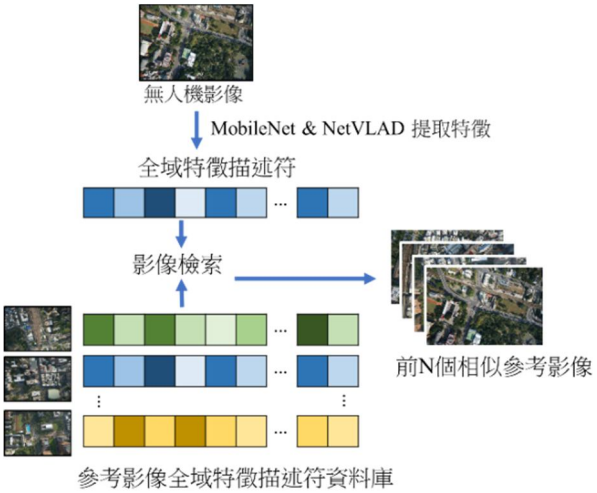


圖 2 影像檢索示意圖

此外，參考影像為帶有地理資訊之正射影像，地面空間解析度 (GSD) 與無人機影像解析度存在明顯差異，以本研究為例前者 0.5m，後者 0.0545m (影像資訊細節請見 3.2 節)，因此匹配前需先調降無人機影像解析度使其影像空間解析度一致化，避免影響匹配成果。本研究改善影像匹配之部分測試與分析，是直接將參考影像裁切，範圍以無人機本身拍攝時紀錄的外方位參數再加上參考影像地面空間解析度推算獲得，根據結果即可裁切出適當範圍之參考影像，同時外方位參數可做後續空間後方交會定位參數初始值。此外，資料部分還需 DSM (請見 3.2 節)，供後續空間後方交會解算。

本研究中之影像特徵點匹配方法使用的深度學習演算法 SuperPoint 以及 SuperGlue，前者以影像為輸入並提取其中特徵點與其描述符，後者輸入特徵點與其描述符並透過模型計算出其中之匹配關係。

2.2 影像特徵點萃取與匹配

本研究中影像特徵點匹配方法使用深度學習演算法 SuperPoint (DeTone *et al.*, 2018) 以及 SuperGlue (Sarlin *et al.*, 2020)，前者以影像為輸入並提取其中特徵點座標與其描述符，後者輸入特徵點座標與其描述符並透過模型計算出其中之匹配關係。

SuperPoint 為基於自監督式學習的特徵點之提取及描述器，可對一張影像提取其中特徵點之影像座標、信心值以及其固定長度之描述符，其架構如圖 3 所示。SuperPoint 之架構由全卷積神經網路構成 (fully-convolutional neural network)，並且

主要分為兩部分。第一部分為單一共享的編碼器(encoder)，將輸入之影像維度降維處理；第二部分包含兩個解碼器 (decoder)，第一個解碼器提取特徵點之影像座標、信心值，第二個解碼器產生特徵點之描述符。第一部分之單一共享的編碼器 (encoder)為一似 VGG 結構之編碼器，由 8 個3×3之卷積層以及 3 個最大池化層 (Max-pooling) 組成。其中卷積層將影像之通道數從 1 提高到 256 維，而最大池化層使原本長寬為$H \times W$之影像縮小為$H/8 \times W/8$。

第二部分包含兩個解碼器 (decoder)，分別輸入第一部份輸出之特徵圖，經過第一個解碼器後提取特徵點之影像座標、信心值 (Interest Point Decoder)。第二個解碼器提取特徵點之描述符(Descriptor Decoder)。

第一個解碼器 Interest Point Decoder 將第一部分輸出之特徵圖通過卷積層後大小從原本 $H/8 \times W/8 \times 128$轉換為 $H/8 \times W/8 \times 65$ 之特徵圖。一個像素單元的通道深度 (channel depth) 從 128 轉換為 65 對應到原始影像8×8像素的區域再加上一個額外的無特徵點放置區 (dustbin)。經過 Softmax 後，去除 dustbin 使其轉換為 $H/8 \times W/8 \times 64$之特徵圖，最後轉換形狀 (reshape)，最後輸出結果為大小為 $H \times W \times 1$，每一個像素單元最多產生一特徵點座標 x、y 以及該點之信心值 c。

第二個解碼器提取特徵點之描述符(Descriptor Decoder)，將第一部分之特徵圖通過卷積層後大小從原本 $H/8 \times W/8 \times 128$轉換為 $H/8 \times W/8 \times D$ (D一般預設為256) 之描述向量，經過雙三次插值 (Bi-cubic interpolation) 以及正規化後(L2-normalizes)，最後輸出結果轉換大小為$H \times W \times D$。

SuperGlue 為一種匹配兩組特徵的神經網路，主要包含兩部分，其架構如圖 4 所示。第一部分為注意力圖像神經網路 (Attentional Graph Neural Network)，其中自注意力機制 (Self-attention) 中目標影像與參考影像上的候選特徵需與自身影像先行比較後，選出最具獨特性的特徵後，接著利用交叉注意力機制 (Cross-attention) 與參考影像上特徵進行比較。

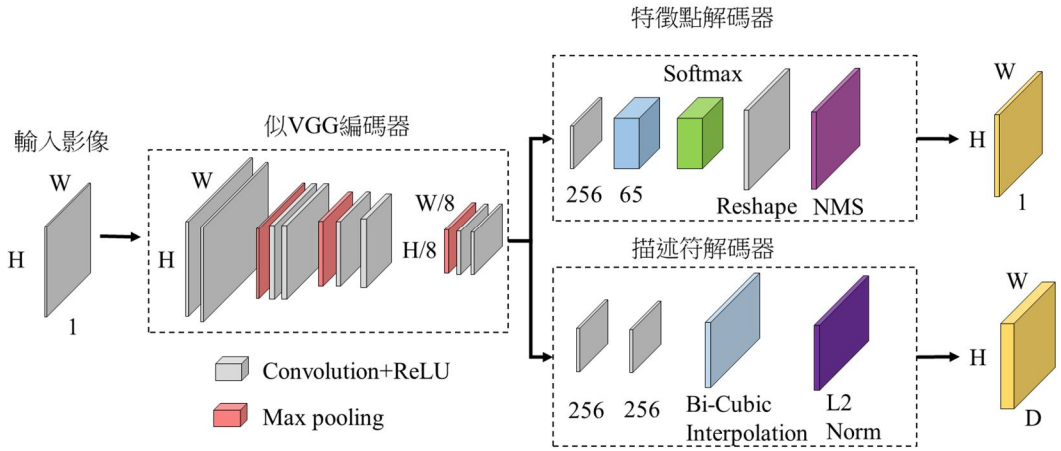


圖 3 SuperPoint 模型架構

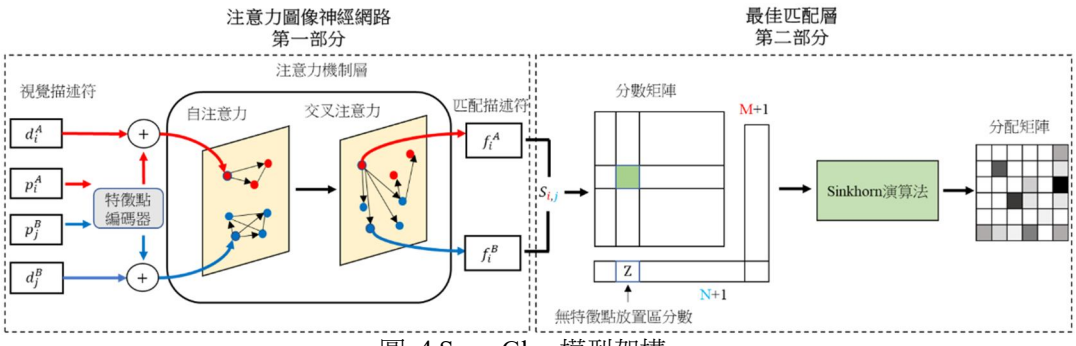


圖 4 SuperGlue 模型架構

注意力圖像神經網路 (Attentional Graph Neural Network) 承接 SuperPoint 最後的輸出，影像中之特徵點位置資訊p_i (包含特徵點之座標 x、y，以及信心值 c)、特徵點描述符d_i (預設為一維度為 256 之向量)，同時匹配兩張圖中之特徵點。如圖 4 所示，$(p_i^A,\ d_i^A)$ 為圖 A 之特徵點 i 位置資訊以及描述符，$(p_j^B,\ d_j^B)$ 為圖 B 之特徵點 j 位置資訊以及描述符。藉由將p_i資訊與d_i資訊整合，特徵點位置資訊p_i先通過關鍵點編碼 (Keypoint Encoder) 提高位置資訊的維度再與描述符d_i相加在一起。而後通過圖像神經網路 (Graph Neural Network, GNN) 及注意力機制分析點之間空間分布關係，以克服重複紋理之場景。注意力機制分為 (1)自注意力機制 (Self-attention)，分析同一張影像中一特徵點與其他特徵點之間的關聯，使相似之各點具有其獨特性；(2)交叉注意力機制 (Cross-attention)，分析一張影像中一特徵點與另一影像特徵點之間的關聯，比對較佳匹配。通過注意力機制後輸入第二部分最優匹配層，計算各點之間內積距離，最後輸出分數預測矩陣 **S**。並透過加入無特徵點放置區 (dustbin) 剔除無法匹配特徵點，如被遮蔽或視線不佳之特徵，以提升匹配之可靠度與穩定性。最後採用 Sinkhorn 演算法迭代數次，得到特徵點之最佳分配矩陣 **P**。

2.3 SuperPoint 描述符與 SIFT 描述符比較

SuperPoint 模型描述符由預設長度 256 的向量組成，表示一個維度 256 的特徵空間，描述圖像中特徵點周圍區域的訊息或數據結構，主要目的是在圖像之間找到相同或相似的特徵點。SIFT 尺度不變特徵變換演算法，通過計算特徵點周圍區域的梯度方向直方圖來生成描述符，對光照和噪聲有一定的抵抗性，對於尺度和旋轉也具有不變性，其特徵描述符由長度 128 的向量組成。

為比較兩種描述符的特性，將特徵向量以視覺化的方式使用極座標圖描繪出來。首先選定一張參考圖以及另一張目標影像如圖 5 所示，使用 SuperPoint 以及 SIFT 分別提取參考圖特徵點 (綠色原點) 座標及描述符；目標影像以 90°旋轉 3 次，於四個方向使用 SuperPoint 以及 SuperSIFT 分別提取參考圖特徵點座標及描述符。選取一同名點特徵，以極座標圖方式將特徵向量視覺描繪，結果如圖 6 所示。經比較後 SuperPoint 描述符除了 0°相對旋轉時參考與目標描述符相似，其餘旋轉角度皆無相似的描述符樣式；SIFT 描述符之參考與目標描述符於所有相對旋轉角度下皆展現相似描述符樣式。由於參考影像與目標影像原本就存在約 20°相對旋轉量，可推論 SuperPoint 描述符於小角度時擁有旋轉不變性，SIFT 描述符則於各種角度皆擁有旋轉不變性。

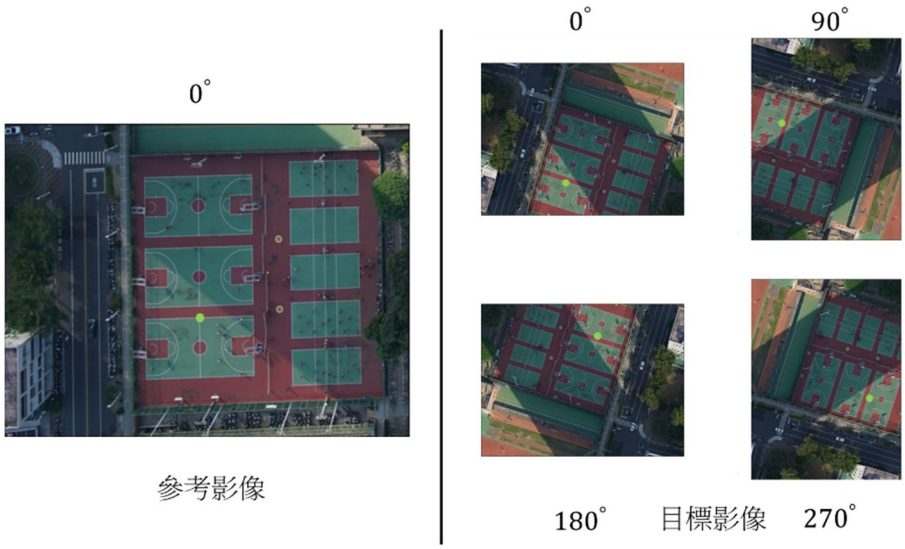


圖 5 左側為參考影像、右側為目標影像旋轉四個方向後結果，圖中角度為圖片旋轉角度

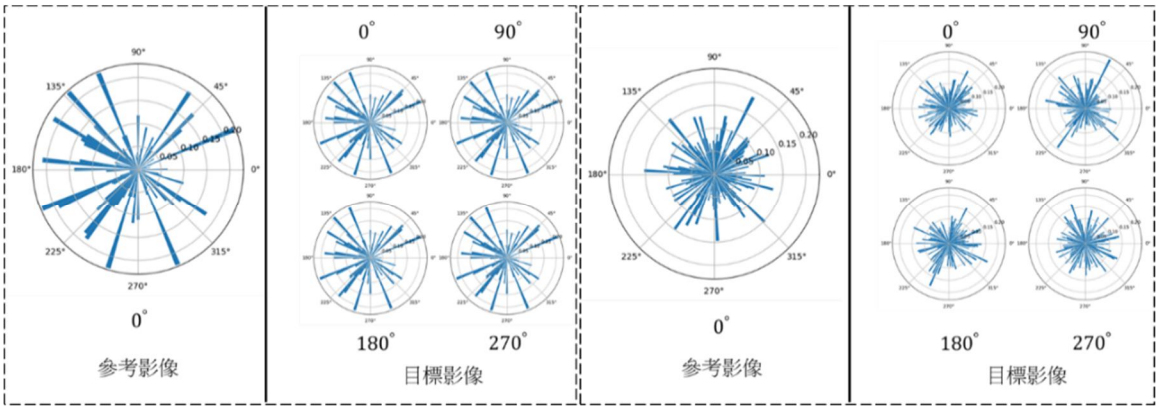


圖 6 左圖為 SuperPoint 描述符，右圖為 SIFT 描述符，圖中角度為圖片旋轉角度

2.4 SuperSIFT 與描述符提升維度

為改善匹配模型對於旋轉的抵抗性，採用 SIFT 描述符取代 SuperPoint 描述符，並取名為 SuperSIFT。由於 SuperPoint 之座標提取和特徵描述符萃取於模型中屬分枝的結構，可不使用特徵描述符分枝結果，於 SuperPoint 提取座標位置，並使用 SIFT 提取該位置描述符。使用 SuperSIFT 特徵訓練 SuperGlue 時，為使得模型收斂更快且匹配成果更好，訓練時採用 SuperGlue 預訓練模型，該模型使用 SuperPoint 特徵進行預訓練。SuperGlue 的結構中，第一部分通過特徵點編碼器，將座標點位置、信心值提升維度使其與描述符維度都為 256 後，相加聚合兩樣資訊再輸入注意力機制層。由於訓練時採用預訓練模型，且輸入特徵點描述符之維度一旦確定模型大小即不能變更，因此使用 SuperSIFT 提取特徵並訓練 SuperGlue 時，長度 128 之 SIFT 特徵點描述符需先經過描述符編碼器提升維度，才能與預訓練模型的大小相匹配，如圖 7 所示。描述符提升維度：本研究提出兩種方法分別為內插法以及可學習參數法。

2.4.1 內插法

SuperSIFT 特徵描述符維度為 128，而 SuperPoint 特徵描述符維度為 256 是 SIFT 特徵描述符兩倍的長度，內插法可用描述符向量中前後項兩兩相加並平均得到內插值，並且最後一項與第一項相加平均得到。圖 8 中數字部分為向量的索引，其中內插後描述符向量最後由 SuperSIFT 描述符最後一個索引與第一個索引內插得到。此種提升維度的方法為非學習參數方法，不需要增加額外的模型參數量，結構較簡單且訓練也較快，內插方法計算容易，且不須微調模型即可使用經過資料擴增的訓練資料改善模型的旋轉抵抗性。然而此方法無法提供更多的特徵點描述資訊，因此本研究再提出另一個方法，並於後續比較與分析。

2.4.2 可學習參數法

可學習參數法則使模型自行學習如何提升 SuperSIFT 特徵描述符維度至與 SuperPoint 特徵描述符維度相同，如圖 9 所示。其結構由兩組可學習參數組成，每組由一層卷積層、一層 Batch Normalization、一層 Tanh 激活函數組成。SuperSIFT 特徵描述符向量值大約介於[0.2~-0.2]，Tanh 激活函數取值介於[-1,1]，因此相較於 Sigmoid、ReLU 等取值後大於 0 之激活函數，較能保留特徵描述符資訊因此更適合模型學習。

此方法須經過採用兩階段式的模型訓練 (如圖 10 所示)，目的是使得模型中的注意力機制層與最優匹配層也能學習適應 SuperSIFT 描述符。第一階段凍結除了提升維度部分以外的其他模型參數，並用較大的學習速度訓練模型，使得提升維度的編碼器部分適應 SuperSIFT 描述符與 SuperPoint 描述符的差別。第二階段模型訓練解除第一階段中凍結的參數後，同樣使用較小的學習速度微調整個模型，使得模型中的注意力機制層與最優匹配層學習適應 SIFT 描述符。

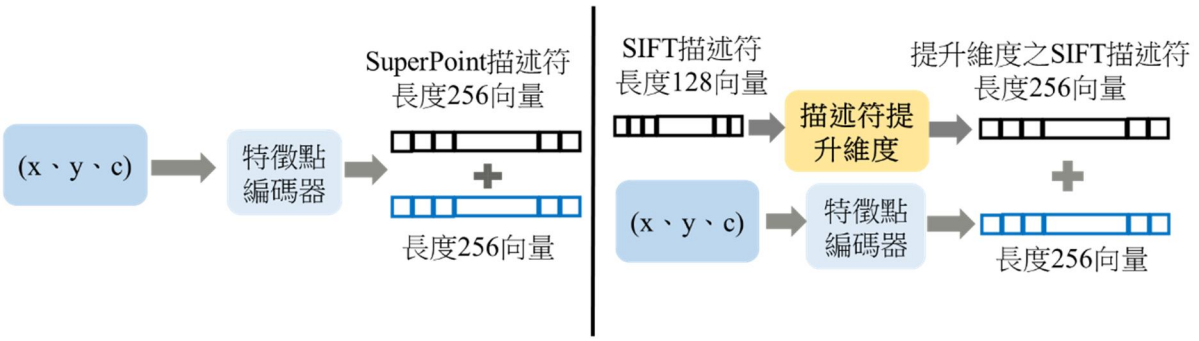


圖 7 左側為 SuperGlue 中原特徵編碼結構，右側為加入描述符提升維度後結構

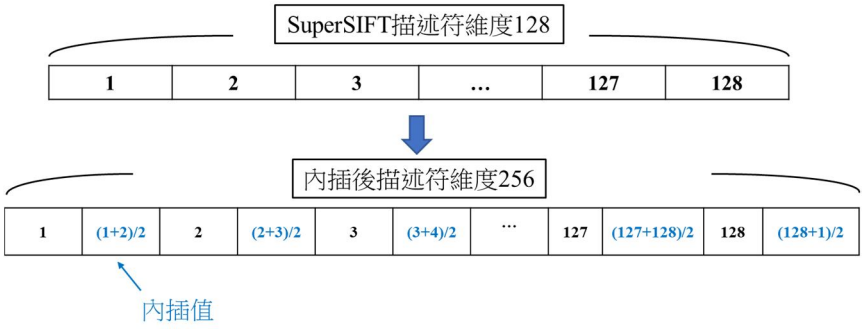


圖 8 內插法計算示意圖

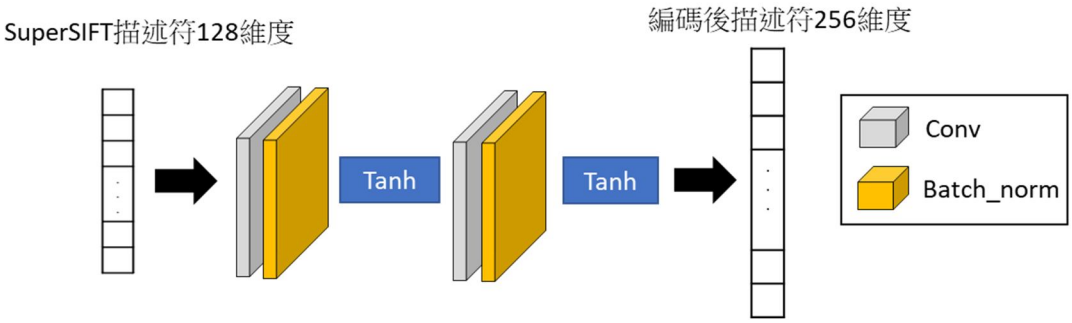


圖 9 可學習參數法計算示意圖

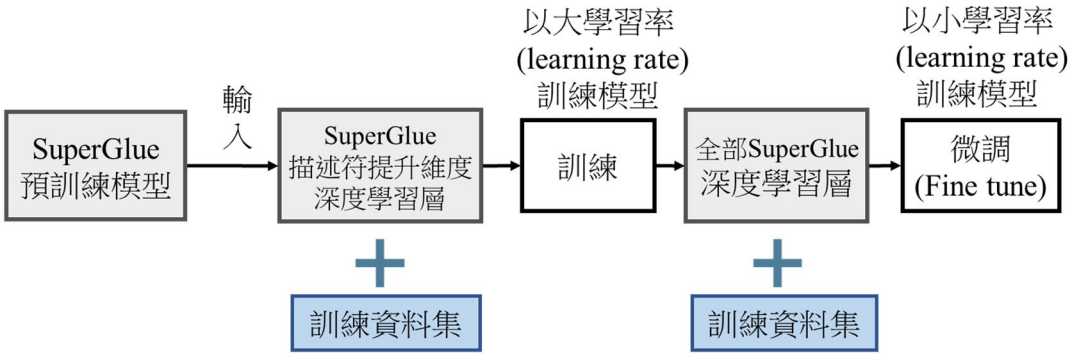


圖 10 兩階段式的可學習參數訓練流程

2.5匹配模型擴增訓練資料集

2.5.1資料擴增

SuperGlue 模型訓練需要多組匹配特徵點對，包含特徵點的座標以及描述符。訓練資料中影像沒有相對旋轉，因此對於模型來說難以或無法推論影像有相對旋轉情形時的匹配。為增加模型對於旋轉的適應性，加入隨機角度的平面旋轉矩陣，以影像為中心旋轉影像，再輸入 SuperPoint 模型產生特徵點，比對後製作擴增訓練資料集。影像隨機旋轉以影像中心為基準，**M** 為影像旋轉之轉換矩陣，則其轉換公式如下：

$$\mathbf{M}=\begin{bmatrix}\alpha & \beta & (1-\alpha)*c_x-\beta*c_y\\ -\beta & \alpha & \beta*c_x+(1-\alpha)*c_y\end{bmatrix} \quad \text{.....(1)}$$

$$\alpha = scale*\cos\theta$$

$$\beta = scale*\sin\theta$$

其中，c_x、c_y為影像的中間點座標，θ為旋轉角度，scale 為縮放比例預設為 1。

2.5.2擴增訓練資料集流程

本研究使用 GL3D 以及 Tourism 影像資料集訓練匹配模型 (細節請見 3.1 節)，其中紀錄每張影像之內、外方位參數，且紀錄相鄰兩影像間之重疊率、深度圖，使得影像對之間相對關係可確立。產生訓練 SuperGlue 模型訓練資料集，為使模型增加對於相對旋轉情境的適應能力，擴增訓練資料集製作步驟如圖 11 所示。

(1) GL3D 成對的影像先經過介於某一區間之隨機旋轉，再使用 SuperPoint 提取每張影像中特徵點資訊，每個點包含其位置、信心度與其描述預設找出信心度最高之 2000 個點。

(2) 訓練 SuperGlue 需輸入成對的特徵點，即兩張影像中的共軛點。由於使用內、外方位參數需要未經過旋轉之影像，使用$\boldsymbol{M^{-1}}$逆旋轉矩陣將經旋轉影像之特徵點逆轉換回無旋轉特徵位置。

(3) 使用 GL3D 資料集提供之內、外方位參數及深度圖資訊，以投影方法比對特徵點，方法如下：現有兩張成對影像 A 和 B，各自有其內方位參數矩陣$\boldsymbol{K_A}$、$\boldsymbol{K_B}$，深度圖Map_A、Map_B，外方位參數紀錄之相機姿態旋轉矩陣$\boldsymbol{R_A}$、$\boldsymbol{R_B}$以及平移矩陣$\boldsymbol{t_A}$、$\boldsymbol{t_B}$。設影像 A 中有一特徵點p_A，特徵點p_A為於影像 A 中在影像座標系下之特徵點座標，先透過內方位參數矩陣$\boldsymbol{K_A}$投影到相機坐標系 (cameracoordinate system)。

$$p_{camera\ coordinate\ system} = \boldsymbol{K_A}^{-1}p_A \quad \text{.....(2)}$$

同時透過影像深度圖Map_A內插獲取特徵點p_A之深度值 z：

$$\mathrm{z} = \mathrm{interpolate\_depth}(p_A(\mathrm{x},\mathrm{y}), Map_A) \quad \text{.....(3)}$$

將特徵點$p_{camera\ coordinate\ system}$添加成齊次座標 $\mathrm{p}_{Homogenous}$：

$$\mathrm{p}_{Homogenous} = [p_{camera\ coordinate\ system}, 1] \quad \text{.....(4)}$$

齊次座標乘以該特徵點深度計算三維座標p_{xyz}：

$$p_{xyz} = \mathrm{p}_{Homogenous}\cdot \mathrm{z} \quad \text{.....(5)}$$

使用旋轉矩陣**dR**以及平移矩陣**dt**進行座標變換：

$$p_{xyz}{}' = \mathbf{dR}\cdot p_{xyz} + \mathbf{dt} \quad \text{.....(6)}$$

其中$\mathbf{dR} = \boldsymbol{R_B}\cdot \boldsymbol{R_A}^{-1}$, $\mathbf{dt} = \boldsymbol{t_B} - \mathbf{dR}\cdot \boldsymbol{t_A}$

將三維座標$\boldsymbol{p_{xyz}}'$投影到影像 B 之相機座標系：

$$\mathrm{p}' = \frac{p_{xy}'}{p_{z}'} \quad \text{.....(7)}$$

將投影特徵點從相機坐標系投影到影像座標系：

$$p_B = \boldsymbol{K_A}\mathrm{p}' \quad \text{.....(8)}$$

依據上述過程將 A 影像特徵點投影到 B 影像，並根據歐基里得距離取一閾值，若小於該閾值則認定兩點為同一點。同時將 B 影像特徵點投影到 A 影像，同樣根據歐基里得距離取一閾值，若小於該閾值則認定兩點為同一點，若確認兩點為彼此共同最鄰近點，即該兩點為一組共軛點。

(4) 根據步驟三比對結果，將步驟(1) 提取之特徵點資料整理並留下共軛點。由於訓練 SuperGlue 模型除須要成對特徵點，也需要無法匹配特徵點，因此除共軛點以外之特徵點也另外記錄起來，以便訓練模型無特徵點放置區中之可學習參數。

SuperGlue 模型訓練分為兩階段式訓練，如圖 12 所示，目的是使訓練過程更容易收斂，以漸進的方式加大相對旋轉角度。

第一階段訓練使用預訓練模型並輸入隨機旋轉角度介於 90° ~ - 90°之資料擴增訓練資料集，以小學習率微調 (Fine tune) 模型。待模型收斂後，再以第一階段訓練之模型為預訓練模型並輸入隨機旋轉角度介於 180° ~ - 180°之資料擴增訓練資料集，以更小之學習率微調模型。兩階段式訓練使模型能漸進的適應旋轉的情形，增加資料多樣性的同時也能避免收斂困難。

2.6 空間後方交會求解外方位

空間後方交會的基本原理是基於光線的直線性質來求解外方位參數，如圖 13 所示。藉由投影中心、像素點和地面點之間形成的共線條件，可以利用至少三個已知地面控制點的座標及其在影像上對應的像點座標列出共線方程式。

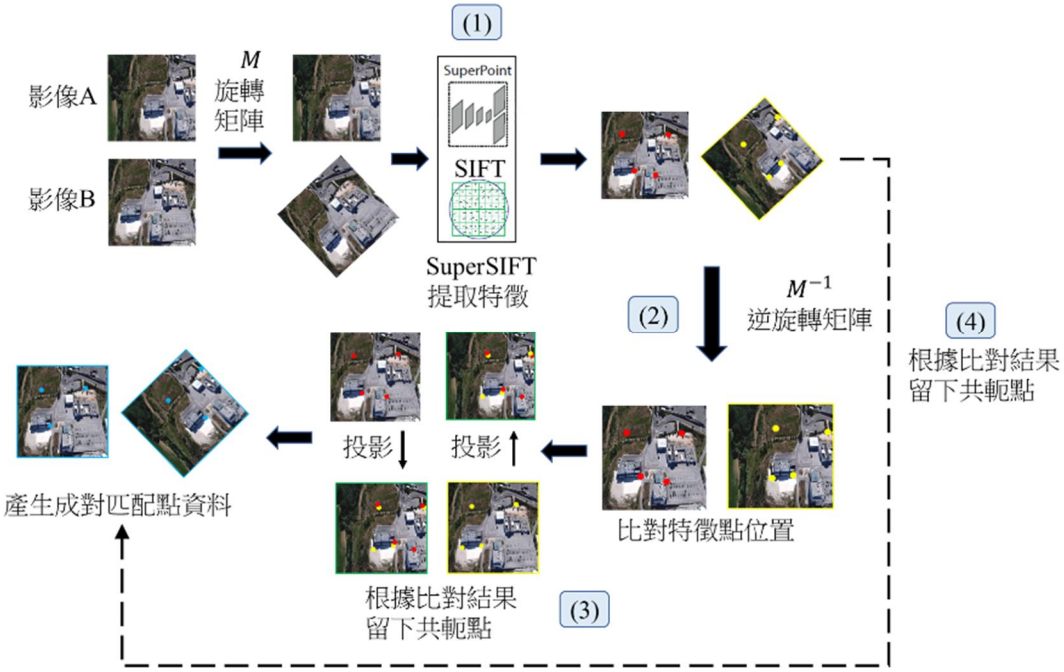


圖 11 完整匹配模型擴增訓練資料集產生流程圖

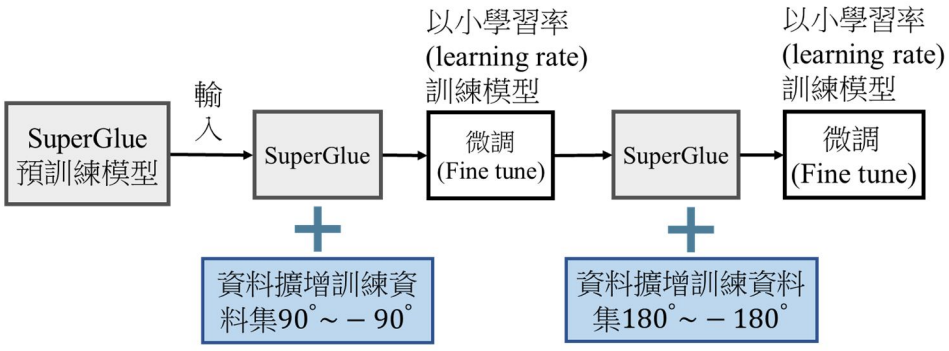


圖 12 以擴增資料集訓練 SuperGlue 模型流程示意圖

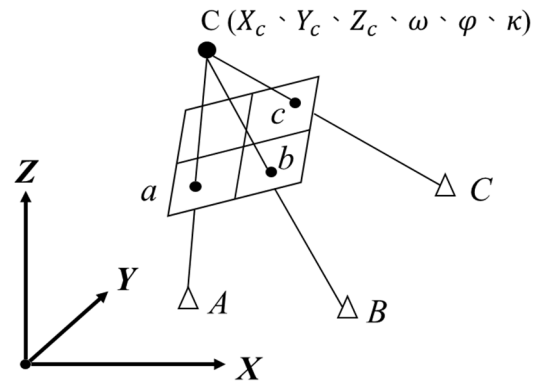


圖 13 單片空間後方交會示意圖

根據共線方程式反求相片的外方位元素，即X_c、Y_c、Z_c、ω、φ、κ。欲求得單張相片的外方位元素，需獲取待求外方位之目標影像與參考影像之間共軛點。參考影像之共軛點座標與地面採樣距離(Ground sample distance)、影像左上角地理座標以及DEM 可計算出共軛點之物空間座標。共軛點像座標與物空間座標可組成共線方程式。共線條件方程式如下：

$$x - x_0 = -f\frac{a_1(X-X_c)+b_1(Y-Y_c)+c_1(Z-Z_c)}{a_3(X-X_c)+b_3(Y-Y_c)+c_3(Z-Z_c)} \quad (9)$$

$$y - y_0 = -f\frac{a_2(X-X_c)+b_2(Y-Y_c)+c_2(Z-Z_c)}{a_3(X-X_c)+b_3(Y-Y_c)+c_3(Z-Z_c)} \quad (10)$$

其中，f為像機焦距；x_0、y_0為像主點；x、y為地面已知點的像座標；X、Y、Z為地面已知點的物空間座標；X_c、Y_c、Z_c為投影中心的物空間座標；a_1、a_2、a_3、b_1、b_2、b_3、c_1、c_2、c_3為相片姿態參數 (ω、φ、k) 組成的旋轉矩陣。當地面點已知，相機外方為元素X_s、Y_s、Z_s、ω、φ、k未知，空間後方交會共線式可以透過泰勒級數展開成線性方程組並表示為矩陣形式如下：

$$\boldsymbol{AX} = \boldsymbol{L} + \boldsymbol{V} \quad (11)$$

當匹配共軛點大於三組，此時觀測數方程式大於未知數數量，會產生多於觀測量，可根據最小二乘法原理求得$\boldsymbol{X}$中各項未知參數改正值如下式：

$$X = (\boldsymbol{A}^{\boldsymbol{T}}\boldsymbol{A})^{-1}(\boldsymbol{A})^{\boldsymbol{T}}\boldsymbol{L} \quad (12)$$

根據預先設定門檻值，迭代計算直到未知參數改正值與上一次迭代之未知參數改正值差值小於門檻值，則迭代計算收斂，最後再加上給定初始外方位便可求得六個外方位參數。

3. 實驗成果

實驗結果與討論主要分五個部分：3.1 SuperGlue 模型訓練與驗證影像資料集、3.2 空間後方交會分析資料集、3.3 單片匹配具相對旋轉比較分析、3.4 無人機影像資料集解算外方位誤差分析、3.5 匹配密度及匹配分散程度分析。

3.1 SuperGlue 模型訓練與驗證影像資料集

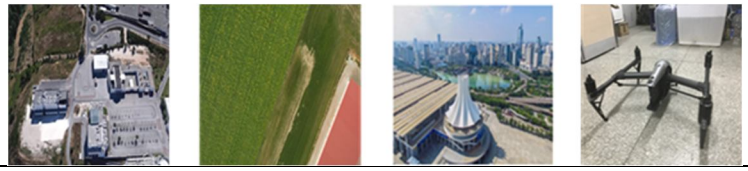
本研究於訓練 SuperGlue 模型時使用由 Shen *et al.* (2018) 提出為完成影像檢索研究而創建之 GL3D 基準資料集。該研究基於批量三元组損失函數結合網格重投影的方法訓練 CNN 模型，所提方法顯著加速 3D 重建中的影像檢索過程，並因此創建 GL3D 資料集訓練深度學習模型，全名 Geometric Learning with 3D Reconstruction，詳細資料如表 1 所示。包含都市、鄉村、旅遊景點、小物件等影像。總共 378 個子資料集中各一種場景，總共 90590 張高解析度影像，每個場景包含 50 到 1000 張影像且交疊比例高，適合應用於幾何相關且 3D 資訊豐富之情境，諸如：特徵軌跡匹配、相機姿態、點雲資料、網格模型。有別於其他資料集 Oxford5k (Philbin *et al.*, 2007)、Paris6k (Philbin *et al.*, 2008) 以及 Holiday (Jegou *et al.*, 2008)，有以下幾點特色。

(1) 與現有資料集最大區別為每個資料集完整覆蓋該場景中的地物，每張影像間密集相連且平均分布於地物周圍的不同位置和視角。

(2) GL3D 表現弱語意，資料集中包含缺乏紋理如草地、河流等場景，於其他資料集中不常見。

(3) 豐富的幾何情境，影像間重疊率高且緊密連結，不僅有兩視角影像對，多視角觀測更好因應精確的幾何運算如相機姿態、點雲、網格模型等。

表 1 GL3D 基準資料集資訊

資料集名稱	Geometric Learning with 3D Reconstruction 資料集
影像類型	空拍 (如：都市、鄉村)　　近景 (如：旅遊景點、小物件)
影像波段	彩色波段 (R、G、B 波段) 產生訓練資料時會轉換至灰階影像
影像大小	影像長寬為 1000*1000
資料集影像張數	90590 張高解析度影像
影像範例	

GL3D 資料集的一個潛在缺點是影像在短時間內拍攝，因此缺乏光照、天氣和季節變化。儘管可以應用光度數據增強，仍應尋求更多真實的數據來改進學習模型。Shen 等人參照 SiaMAC (Radenović *et al.*, 2016)，也從網路公共旅遊數據集中生成了幾何標籤，以進一步增加數據的多樣性。具體來說，從網路公共旅遊數據集中下載並提取影像，然後通過 3D 引擎重建每個數據，最終獲得了被認為構建良好的 530 個場景 (55,657 張影像)，並命名為 Tourism 基準資料集，詳細資料如表 2 所示。所有場景皆是旅遊景點，包含從多種角度拍攝之建築物的外觀，且於不同時間拍攝，可看出有明顯光照、天氣和季節變化。經由 Tourism 基準資料集訓練後的模型可更好的適應不同時間點光照、天氣和季節變化。

3.2 空間後方交會分析資料集影像資料集

成大校園無人機空拍圖總共 839 張影像，影像大小長寬為 7952×5304，地面空間解析度為 0.0545 m，無人機飛行高度 221 m，使用 sony Alpha 7R II 全片幅相機，選定焦距 15 mm 之鏡頭，其可視範圍 (Field of view) 可達 110°。各影像之外方位參數使用 Metashape 攝影測量軟體，進行影像空三平差得到六個外方位參數 (X、Y、Z、ω、ϕ、κ) 作為真值供後續空間後方交會解算分析誤差。其中地面控制點誤差如表 3 所示，平均無人機位置誤差如表 4 所示。

空間後方交會時主要利用參考影像匹配共軛點平面座標。如圖 14 所示，參考影像大小長寬為 3114×2787，為地面空間解析度 0.5 m 之正射影像，以 Metashape 攝影測量軟體將多張無人機影像合成製作，解算時還需搭配數值地表模型獲得高程。

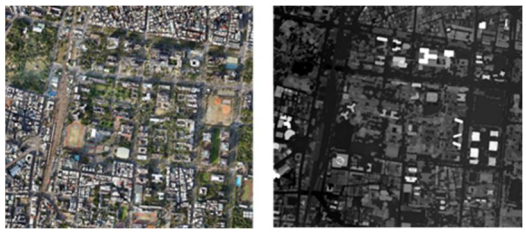


圖 14 成大校園參考影像 (左) 與 DSM (右)

數值地表模型或稱 Digital surface model (DSM)，是一種數位化的地理數據表示形式，用於描繪地球表面上所有可見物體的高度，包括建築物、植被和其他構造物，用途是取得匹配特徵點於物空間座標高程值，其製作同樣由 Metashape 攝影測量軟體執行，空間解析度為 0.25 m。

3.3 單像匹配具相對旋轉結果評估與比較分析

本研究包含傳統影像特徵匹配方法 SIFT+FLANN、SUFT+FLANN，兩者皆加入 RANSAC (Fischler *et al.*, 1981) 濾除粗差。深度學習影像特徵匹配方法有 SuperPoint+ SuperGlue 與本研究提出之 SuperSIFT (內插法) + SuperGlue (擴增) 以及 SuperSIFT (可學習參數法) +SuperGlue (擴增)，由於深度學習方法本身包含濾除粗差能力 (2.2 節提到之無特徵點放置區)，因此未加入 RANSAC。

表 2 Tourism 基準資料集資訊

資料集名稱	Tourism 資料集
影像類型	近景，如：旅遊景點、建築物
影像波段	彩色波段 (R、G、B 波段) 產生訓練資料時會轉換至灰階影像
影像大小	影像長寬為 1000*1000
資料集影像張數	55,657 張高解析度影像
影像範例	

表 3 地面控制點誤差

點數	X 誤差 (cm)	Y 誤差 (cm)	Z 誤差 (cm)	XY 誤差 (cm)	總誤差 (cm)
14	11.1498	12.5868	5.91888	16.815	17.8263

表 4 平均無人機位置誤差

X 誤差 (m)	Y 誤差 (m)	Z 誤差 (m)	XY 誤差 (m)	總誤差 (m)
0.112799	0.138552	0.992594	0.178662	1.00855

前述總共五種匹配方法分別對測試資料集匹配，每匹配完一次逐漸旋轉目標影像每次 30°，共 0°、30°、60°、90°、120°、150°、180°五種角度並分別與參考影像匹配，比較匹配點對之數量，其中一匹配影象對中左圖為參考影像，右圖為目標影像，結果如圖 16~圖 21 所示。

圖 15 為兩張拍攝位置不同之垂直空拍成對影像，其中包含高樓、一般建物、城市景觀，左圖為參考影像，右圖為目標影像。

根據表 5 比較的結果可分析出，由於 SIFT 與 SURF 萃取特徵本身具有旋轉不變性，因此於各角度都能得到約 100 個匹配特徵點對。SuperPoint+SuperGlue 之匹配方法則是可應付相對旋轉角度 60°以下之匹配，最多可獲得約 800 個匹配對，然而大於 60°則無法產生匹配對。本研究提出第一個方法 SuperSIFT (內插法) + SuperGlue (擴增資料集)，於 90°以下可獲得約 400 個匹配對，於 90°以上則可匹配得約 300 個匹配對，180°則有 133 個匹配。本研究提出之第二個方法 SuperSIFT (可學習參數法) + SuperGlue (擴增資料集)，除 180°於各角度都能獲得約 400 個以上之匹配對。相比之下本文提出之方法可在各角度下皆獲得足夠之匹配點數量供後續解算外方位，且匹配數量相較傳統方法 SIFT+FLANN 與 SURF+FLANN 能獲得較多匹配對。此外，SuperPoint + SuperGlue 之匹配方法於相對旋轉角度小時可獲得大量匹配點對，然而當角度大於 60°時則無法產生足夠數量的匹配點對解算外方位。

圖 15 參考影像 (左圖)，目標影像 (右圖)

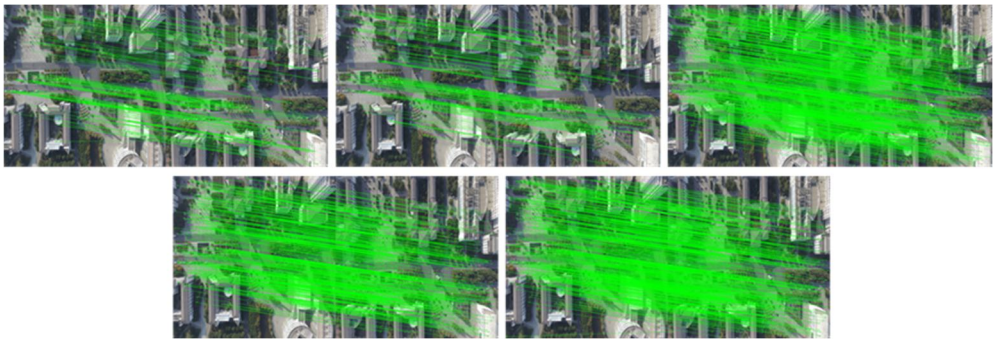


圖 16 城市垂直空拍影像旋轉角度 0°匹配成果，左上、中上、右上、左下、右下依序使用匹配方法 SIFT+FLANN、SURF+FLANN、SuperPoint+SuperGlue、本研究提出 SuperSIFT (內插法) +SuperGlue (擴增資料集)、本研究提出 SuperSIFT (可學習參數法) +SuperGlue (擴增資料集)

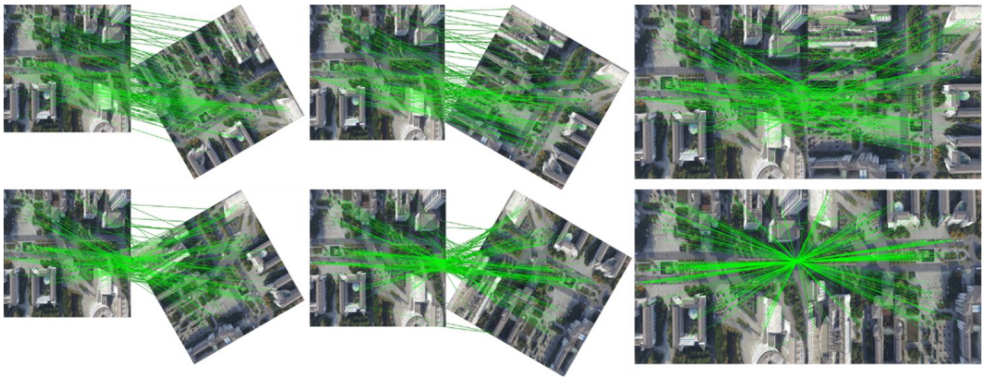


圖 17 城市垂直空拍影像 SIFT+FLANN 方法 左上、中上、右上、左下、中下、右下依序是旋轉角度 30°、60°、90°、120°、150°、180°

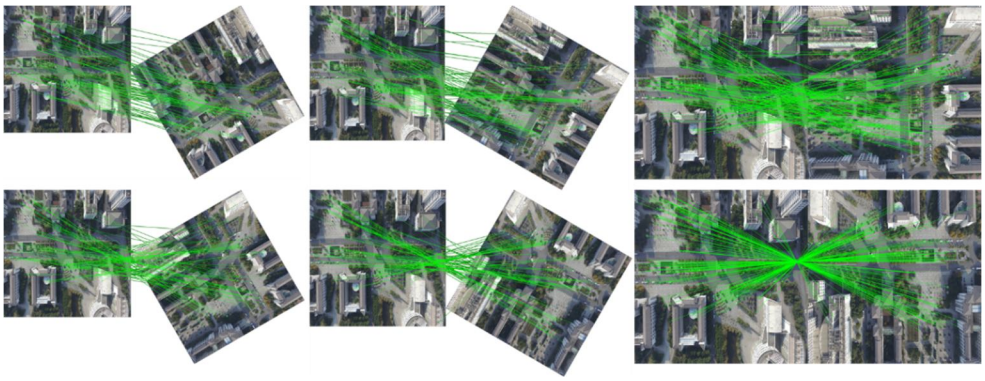


圖 18 城市垂直空拍影像 SURF+FLANN 方法 左上、中上、右上、左下、中下、右下依序是旋轉角度 30°、60°、90°、120°、150°、180°

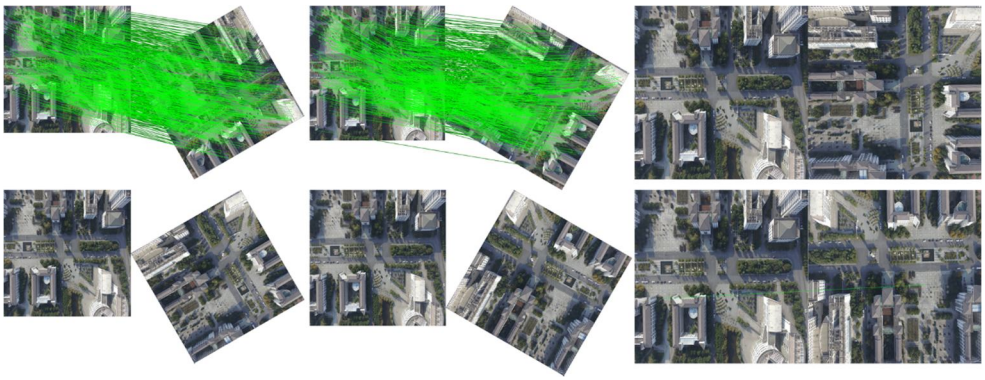


圖 19 城市垂直空拍影像 SuperPoint+SuperGlue 方法 左上、中上、右上、左下、中下、右下依序是旋轉角度 30°、60°、90°、120°、150°、180°

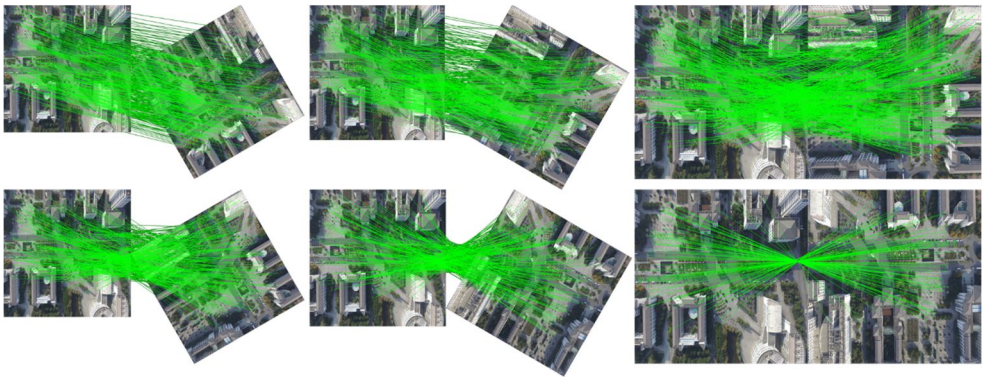


圖 20 城市垂直空拍影像本研究提出 SuperSIFT(內插法)+SuperGlue(擴增資料集) 左上、中上、右上、左下、中下、右下依序是旋轉角度 30°、60°、90°、120°、150°、180°

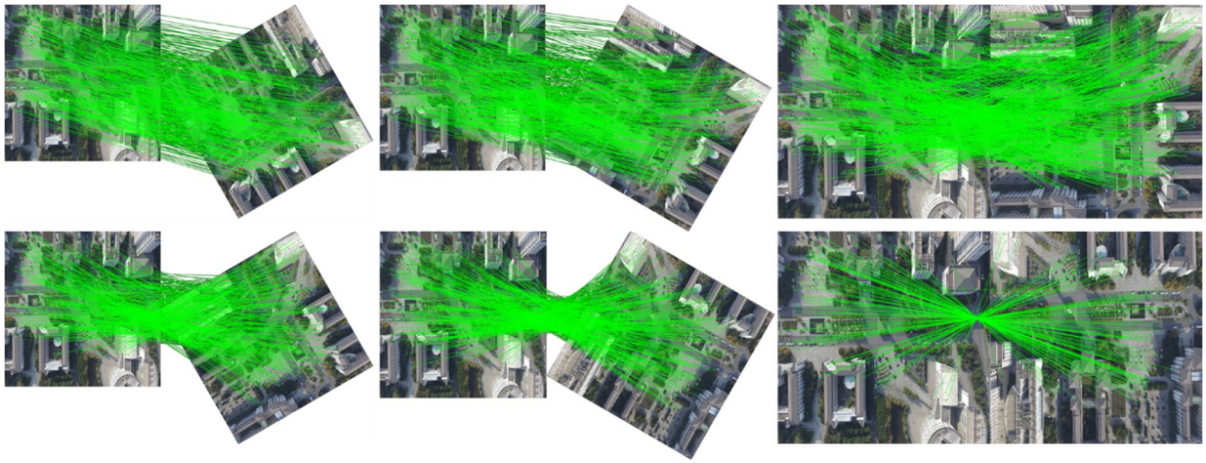


圖 21 城市垂直空拍影像本研究提出 SuperSIFT (可學習參數法) +SuperGlue (擴增資料集) 左上、中上、右上、左下、中下、右下依序是旋轉角度 30°、60°、90°、120°、150°、180°

表 5 各方法匹配點數

匹配方法	目標影像旋轉角度						
	0°	30°	60°	90°	120°	150°	180°
SIFT+FLANN	184	160	154	177	158	158	188
SURF+FLANN	170	90	91	165	96	90	165
SuperPoint+SuperGlue	801	673	637	0	0	0	1
本研究提出 SuperSIFT (內插法)+SuperGlue (擴增)	566	423	425	547	381	327	133
本研究提出 SuperSIFT (可學習參數法)+ SuperGlue (擴增)	759	598	595	664	508	488	157

3.4 無人機影像資料集解算外方位誤差評估與分析

為模擬分析真實案例，以 3.2 節提到 839 張成大校園無人機空拍圖檢測各方法平均匹配點數以及外方位推算誤差，其中誤差計算方式如式 13 所示。其中 839 張影像隨機以相對參考影像 0°、90°、180°、270°旋轉後匹配特徵點，計算平均每張獲得多少匹配點，並統計解算後之外方位誤差。其結果如下，其中誤差以 RMSE 計算：

$$\text{RMSE} = \sqrt{\frac{\Sigma(\hat{\chi}-\chi)^2}{n}} \quad \text{.....(13)}$$

其中$\hat{\chi}$為經最小二乘原理解算得之六個外方位參數 X、Y、Z、ω、ϕ、κ，而χ為已知六個外方位參數。結果如表 6 所示，由其分析得知，傳統匹配方法 SIFT+FLANN 與 SURF+ FLANN 可獲得平均點數相較深度學習法少，SuperPoint+SuperGlue 匹配方

法隨然於大旋轉角度難以匹配，然而小角度時可獲得大量匹配點，因此平均後依然有約 100 個匹配點。本研究提出之 SuperSIFT (內插法)+SuperGlue (擴增資料集) 方法，可獲得平均匹配點數 107 個，SuperSIFT (可學習參數法) +SuperGlue (擴增資料集) 可獲得約 161 個匹配點，皆優於其他方法。外方位解算部分，SIFT+FLANN 與 SURF+FLANN 姿態角度誤差約 10~80 m，位置誤差最大達 20°，由於匹配點個數較少，雖於各方向皆能獲得匹配點，仍相對缺乏多餘觀測量，且其本身也可能因匹配錯誤，導致解算成果較差。SuperPoint+SuperGlue 匹配方法位置誤差約 40 m，姿態角度誤差最大達 37°，較大誤差主要原因是旋轉角度大時，存在只有零星匹配之情形，此時若其中有匹配錯誤點對，會造成解算誤差極大之情形。本文提出之兩種匹配方法，於各旋轉方向皆能獲得足夠數量之匹配點對，因此解算得誤差較小且相對穩定。

如圖 22 為外方位誤差分布圖，顯示出所有 839 張成大校園無人機空拍圖每張影像之外方位誤差分布。根據結果所示，本文提出之兩種匹配方法經解算所得之外方位誤差於三倍標準差外之離群值個數皆小於傳統影像匹配法，而原始 SuperPoint +SuperGlue 匹配方法則由於相對大角度旋轉下幾乎無法匹配，因此離群值相對其他方法最多。由以上所述顯示本文提出之兩種匹配方法相對其他三種方法於有相對旋轉之情形下解算外方位參數更準確且更穩定。

表 6 無人機外方位參數誤差 RMSE

匹配方法	平均分配點數	外方位推算誤差 RMSE (角度單位：度，位置單位：m)					
		ω	ϕ	κ	X	Y	Z
SIFT+FLANN	57.048	6.991	10.453	8.994	20.071	19.880	17.965
SURF+FLANN	23.875	17.413	79.561	6.606	16.408	17.896	13.539
SuperPoint+SuperGlue	104.263	19.071	24.413	37.875	45.022	35.356	47.152
SuperSIFT (內插法) +SuperGlue (擴增)	107.384	3.005	0.846	3.363	4.844	5.936	8.481
SuperSIFT (可學習參數法) +SuperGlue (擴增)	161.654	1.765	1.639	3.278	5.932	6.351	3.600

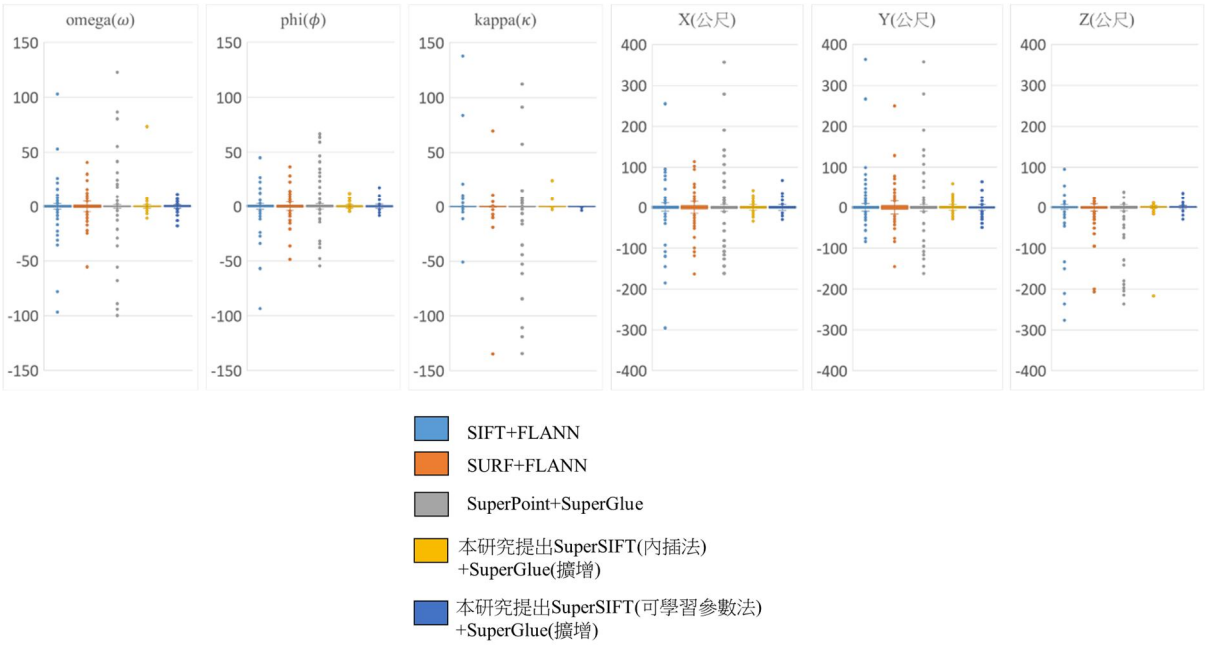


圖 22 成大校園外方位誤差分布圖

3.5匹配密度及分散程度分析

解算外方位時若能盡量滿足以下兩種條件，可獲得較穩定之解算成果。

(1) 匹配點數量越多越好，由於使用最小二乘平差法求解未知參數，越多觀測量可獲得越接近真值之解算。

(2) 匹配點之分布應盡量分散且佈滿整張影像，使得解算時不會產生外插的情況，造成誤差較大成果不穩定。

匹配密度及匹配分散程度計算之方法如下：將影像分成$10 \times 10 = 100$個子區域，並計算總匹配點數與有匹配點之子區域個數。匹配密度為匹配點數除以多少個子區中有匹配點；匹配分散程度則為 100 個子區域中有多少個子區域有匹配點，即整張圖有匹配點的比例。以圖 23 為例，共有 15 個匹配點且 11 個子區域中有匹配點，則匹配密度為$\frac{15}{11} = 1.364$，匹配分散程度為$\frac{11}{100} = 0.11$。

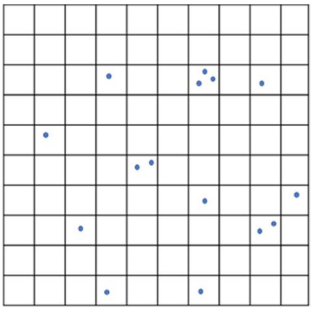


圖 23 子區域劃分及匹配點分布示意圖

以驗證資料集分析匹配密度及匹配分散程度，其中參考影像保持原樣，而目標影像給予隨機旋轉量，經匹配後分別計算匹配密度及匹配分散程度，並將成果平均計算得結果如表 7 所示。

傳統方法部分，SIFT+FLANN 與 SURF+FLANN 方法平均點密度皆較深度學習方法低，平均點分散程度則是前者較高，後者則是所有方法最低。深度學習方法部分，SuperPoint+SuperGlue 方法平均點密度較本文提出之方法低，其平均點分散程度也相對較低，原因是該方法只能應付旋轉角度60°以下之匹配，所以經平均後所得到之各項數據皆較低。本研究提出方法 SuperSIFT（內插法）+ SuperGlue（擴增）與 SuperSIFT (可學習參數法) + SuperGlue (擴增) 方法之平均點密度較其他方法高，其平均點分散程度也較高。顯示出其可獲得較多匹配點，雖有些區塊可能匹配點較集中，然而擁有匹配點之區塊也較多，匹配點分布相較其他方法分散於整張影像各處，有利於後續空間後方交會求解外方位。

表 7 平均點密度及點分散程度比較表

匹配方法	平均點密度	平均點分散程度
SIFT+FLANN	2.012	0.201
SURF+FLANN	1.513	0.098
SuperPoint+SuperGlue	3.162	0.178
SuperSIFT(內插法) +SuperGlue(擴增)	3.266	0.247
SuperSIFT(可學習參數法) +SuperGlue(擴增)	4.247	0.290

4. 結論

本研究旨在提出影像視覺地形輔助定位技術之自動化流程，並將深度學習應用於其中。完整流程首先輸入無人機拍攝影像並與影像檢索搜尋之參考正射影像匹配共軛特徵點。匹配過程應用深度學習模型，由於無人機在航拍過程中常會遇到影像平面旋轉問題，本研究透過增強深度學習模型對旋轉的抵抗能力，使影像視覺地形輔助定位技術更加高效且穩定。針對問題，本文首先提出以 SIFT 描述符取代原有 SuperPoint 描述符，使特徵點具備旋轉不變性，加強了匹配模型 SuperGlue 對旋轉的適應性。同時為加強模型對於旋轉適應能力，提出擴增訓練及驗證資料集的流程，隨機以某一區間角度旋轉影像，增加旋轉多樣性，製作成對特徵點資料集以供 SuperGlue 訓練，改善模型對旋轉問題的適應能力。由於 SIFT 描述符與 SuperPoint 描述符本身結構差異，因此提出兩種可改變 SIFT 描述符使其符合可輸入 SuperGlue 模型結構的方法，第一個方法內插法通過在特徵描述符階段內插 SIFT 特徵

點描述符，增加描述符維度；可學習參數法利用兩組可學習參數取代內插法，同樣增加描述符維度。由以上方法整合出兩種訓練 SuperGlue 模型流程，SuperSIFT (內插法) +SuperGlue (擴增資料集) 以及 SuperSIFT (可學習參數法) + SuperGlue (擴增資料集)，兩種方法皆顯著提高了影像匹配的精度和穩定性。使得影像在不同旋轉角度下均能獲得較多的匹配點對。與傳統的 SIFT+FLANN 和 SURF+FLANN 方法相比，本研究提出的方法在匹配點數量和分散性上具有明顯優勢。雖然 SuperPoint+SuperGlue 方法在相對旋轉角度小於 60°時表現優異，能獲得匹配數量最多匹配點對，但在旋轉角度大於 60°時無法產生足夠的匹配點對。而本研究的兩種方法在所有旋轉角度下均能保持穩定的匹配效果，提高了影像匹配的精度和穩定性。根據本研究提出方法匹配後結果加上 DSM，列出共線式後，以空間後方交會解算外方位平面位置誤差最佳可達 3 m、姿態角誤差最佳可達 1.3°。

參考文獻

黃敬群、黃偉立，2012。無人飛行監控器之定位技術探討，100-101 年度獎勵科技大學及技術學院典範科大計畫產學及研究成果轉專題製作教材，國立高雄應用科技大學。[Huang, C.C., and Huang, W.L., 2012. A study on positioning technology for unmanned aerial surveillance vehicles, 2012-2013 Awards for University of Science and Technology and Technical College, National Kaohsiung University of Applied Sciences. (in Chinese)]

Arandjelovic, R., Gronat, P., Torii, A., Pajdla, T., and Sivic, J., 2016. NetVLAD: CNN architecture for weakly supervised place recognition, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5297-5307, DOI: 10.1109/CVPR.2016.572.

Bay, H., Tuytelaars, T., and Van Gool, L., 2006. Surf: Speeded up robust features, Proceedings of the European Conference on Computer Vision (ECCV), Springer Berlin Heidelberg, pp. 404-417, DOI: 10.1007/11744023_32.

Chang, Y., Ballotta, L., and Carlone, L., 2023. D-Lite: Navigation-oriented compression of 3D scene graphs for multi-robot collaboration, IEEE Robotics and Automation Letters, 8(11):7527-7534, DOI: 10.1109/LRA.2023.3320011.

Chen, H., Luo, Z., Zhang, J., Zhou, L., Bai, X., Hu, Z., Tai, C.L, and Quan, L., 2021b. Learning to match features with seeded graph matching network, Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6281-6290, DOI: 10.1109/ICCV48922.2021.00624.

Chen, S., Wu, X., Mueller, M.W., and Sreenath, K., 2021a. Real-time geo-localization using satellite imagery and topography for unmanned aerial vehicles, Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 2275-2281, DOI: 10.1109/IROS51168.2021.9636705.

DeTone, D., Malisiewicz, T., and Rabinovich, A., 2018. Superpoint: Self-supervised interest point detection and description, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 337-349, DOI: 10.1109/CVPRW.2018.00060.

Fischler, M.A., and Bolles, R.C., 1981. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography, Communications of the ACM, 24(6): 381-395, DOI: 10.1145/358669.358692.

Jegou, H., Douze, M., and Schmid, C., 2008. Hamming embedding and weak geometric consistency for large scale image search, Proceedings of the European Conference on Computer Vision (ECCV), Springer Berlin Heidelberg, Vol.5302, pp. 304-317, DOI: 10.1007/978-3-540-88682-2_24.

Ju, C., Luo, Q., and Yan, X. , 2020. Path planning using

an improved a-star algorithm, Proceedings of the International Conference on Prognostics and System Health Management (PHM-2020 Jinan), pp. 23-26, DOI: 10.1109/PHM-Jinan48558.2020.00012.

Lin, Y., and Medioni, G. , 2007. Map-enhanced UAV image sequence registration and synchronization of multiple image sequences, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, MN, USA, DOI: 10.1109/CVPR.2007.383428.

Lindenberger, P., Sarlin, P.E., and Pollefeys, M., 2023. Lightglue: Local feature matching at light speed, Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, pp. 17581-17592, DOI: 10.1109/ICCV51070.2023.01616.

Low, D.G., 2004. Distinctive image features from scale-invariant keypoints, International Journal of Computer Vision, 60(2): 91-110, DOI: 10.1023/B:VISI.0000029664.99615.94.

Luo, M., Hou, X., and Yang, J., 2020. Surface optimal path planning using an extended Dijkstra algorithm, IEEE Access, 8: 147827-147838, DOI: 10.1109/ACCESS.2020.3015976.

Ma, J., Jiang, X., Fan, A., Jiang, J., and Yan, J., 2021. Image matching from handcrafted to deep features: A survey, International Journal of Computer Vision, 129(1): 23-79, DOI: 10.1007/s11263-020-01359-2.

Muja, M., and Lowe, D.G., 2009. Fast approximate nearest neighbors with automatic algorithm configuration, Proceedings of the International Conference on Computer Vision Theory and Applications (VISAPP), Lisboa, Portugal, Vol.1, pp. 331-340, DOI: 10.5220/0001787803310340.

Philbin, J., Chum, O., Isard, M., Sivic, J., and Zisserman, A., 2007. Object retrieval with large vocabularies and fast spatial matching, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, MN, USA, DOI: 10.1109/CVPR.2007.383172.

Philbin, J., Chum, O., Isard, M., Sivic, J., and Zisserman, A., 2008. Lost in quantization: Improving particular object retrieval in large scale image databases, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, AK, USA, DOI: 10.1109/CVPR.2008.4587635.

Radenović, F., Tolias, G., and Chum, O., 2016. CNN image retrieval learns from BoW: Unsupervised fine-tuning with hard examples, Proceedings of the Computer Vision – ECCV 2016 (Lecture Notes in Computer Science), vol. 9905, pp. 3-20, Springer International Publishing, DOI: 10.1007/978-3-319-46448-0_1.

Rocco, I., Arandjelović, R., and Sivic, J., 2020b. Efficient neighbourhood consensus networks via submanifold sparse convolutions, Proceedings of the Computer Vision–ECCV 2020 (Lecture Notes in Computer Science), vol. 12354, pp. 605-621, DOI: 10.1007/978-3-030-58545-7_35.

Rocco, I., Cimpoi, M., Arandjelović, R., Torii, A., Pajdla, T., and Sivic, J., 2020a. NCNet: Neighbourhood consensus networks for estimating image correspondences, IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(2): 1020-1034, DOI: 10.1109/TPAMI.2020.3016711.

Sarlin, P.E., DeTone, D., Malisiewicz, T., and Rabinovich, A., 2020. Superglue: Learning feature matching with graph neural networks, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4938-4947, DOI: 10.1109/CVPR42600.2020.00499.

Shen, T., Luo, Z., Zhou, L., Zhang, R., Zhu, S., Fang,

T., and Quan, L., 2018. Matchable image retrieval by learning from surface reconstruction, Proceedings of the Asian conference on computer vision, vol.11361, pp. 415-431, Cham: Springer International Publishing, DOI: 10.1007/978-3-030-20887-5_26.

Sinha, D., and El-Sharkawy, M., 2019. Thin mobilenet: An enhanced mobilenet architecture, Proceedings of the 2019 IEEE 10th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), pp. 0280-0285, DOI: 10.1109/UEMCON47517.2019.8993089.

Sun, J., Shen, Z., Wang, Y., Bao, H., and Zhou, X., 2021. LoFTR: Detector-free local feature matching with transformers, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8922-8931, DOI: 10.1109/CVPR46437.2021.00881.

Verdie, Y., Yi, K., Fua, P., and Lepetit, V., 2015. Tilde: A temporally invariant learned detector, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, pp. 5279-5288, DOI: 10.1109/CVPR.2015.7299165.

Yi, K.M., Trulls, E., Lepetit, V., and Fua, P., 2016. Lift: Learned invariant feature transform, Proceedings of the Computer Vision–ECCV 2016 (Lecture Notes in Computer Science), vol. 9910, pp. 467-483, Springer, Cham, DOI: 10.1007/978-3-319-46466-4_28.

Deep Learning-based Image Feature Matching for UAV Visual Positioning

Lai-Han Zou [1*] Chao-Hung Lin [2]

Abstract

When the positioning and orientation equipment on an unmanned aerial vehicle (UAV) is unavailable, visual positioning technology can be utilized to perform spatial resection using only conjugate points from images to derive the vehicle's exterior orientation. This study proposes a visual positioning workflow and addresses the issue of significantly reduced matching success rates when using deep learning models for feature point matching due to planar rotation between images. By incorporating data augmentation with random image rotations, feature points are extracted using a feature extraction model and then input into the matching model for learning. Additionally, the study introduces interpolation methods and learnable parameter methods to replace the feature descriptors used for matching with traditional feature descriptors, enhancing rotational invariance. After extracting and matching feature points from the images, conventional photogrammetry spatial resection can be used to solve for the six exterior orientation elements of the camera mounted on the vehicle, thus achieving vehicle positioning. With the proposed visual positioning workflow, the best achievable plane position error is 3 meters, and the best achievable attitude angle error is 1.3°.

Keywords: Deep Learning, Feature Extraction, Image Matching, Visual Positioning, Rotational Invariance

[1] Master, Department of Geomatics, National Cheng Kung University
[2] Professor, Department of Geomatics, National Cheng Kung University
[*] Corresponding Author, E-mail: ha095863@gmail.com

Received Date: Sep. 16, 2024
Revised Date: Oct. 08, 2024
Accepted Date: Oct.21, 2024