DOI: 10.6653/MoCICHE.201810_45(5).0007

尖端人工智慧於土木設施應用之展望

顏　愉／國立臺灣大學土木工程學系 研究助理
溫欣儀／國立臺灣大學土木工程學系 研究助理
黃尹男／國立臺灣大學土木工程學系 副教授
陳柏華／國立臺灣大學土木工程學系 副教授

時至今日，於建築、工程與營建（Architecture, Engineering, Construction）及設施管理（facilities management）領域中，仍有許多與生活密切相關之問題待解決，諸如應用物件偵測（object detection）、物件追蹤（object tracking）及即時定位與地圖構建（simultaneously localization and mapping, SLAM）技術等等，在解決此類問題的同時，亦須賴以可靠之技術以有效解決問題。目前機器學習（machine learning）技術因其快速有效之特色，成為近年之熱門技術，並且仍高速發展及應用於各項領域。土木及水利工程領域亦逐漸重視此技術，本研究回顧近期於物件偵測、物件追蹤、即時定位與地圖構建、深度估計與影像風格轉換（image-to-image translation）之新穎技術，將上述技術進行初步運用，並提出與土木及水利工程相關應用之建議，作為未來建築、工程與營建及設施管理之參考依據。

緒論

本研究主要針對 2017 年至 2018 年 CVPR（IEEE Conference on Computer Vision and Pattern Recognition）研討會之最新研究，回顧近期於物件偵測、物件追蹤、即時定位與地圖構建、深度估計與影像風格轉換之新穎技術。CVPR 係由 IEEE 每年舉辦之研討會，其所接受的文章皆為電腦視覺領域最具學術創新及應用效益之研究。

目前土木及水利領域，影像處理技術已不可或缺，諸如交通領域之行人追蹤、土木與營建管理領域之影像補遺以及水利工程領域土石流空拍影像處理等等，故本研究將所回顧之文章類型分為偵測與追蹤、即時定位與地圖構建與影像風格轉換。其中，偵測與追蹤再細分為物件偵測及追蹤等二大主題。此外，即時定位與地圖構建再加入影像深度之估計技術加以說明。最後，本研究於回顧機器學習近期技術的同時，提出各項技術於土木及水利領域可能之應用建議。

物件偵測與追蹤

物件偵測

物件偵測係現今數位影像處理（image processing）技術的一主要課題。相較於單純的影像分類（image classification）技術，僅負責針對單一影像進行分類；物件偵測技術需於含有不定個數、不定種類物件的影像中，偵測特定種類之物件（如行人、車輛、建物…等），並獲取其相關資訊，包含被偵測物件之位置及範圍、大小以及種類等。

目前物件偵測技術大量引入深度學習（deep learning）技術，透過複雜且眾多層次的神經網路模式，使得其相較傳統機器學習技術能處理較複雜的課題，在偵測上也較不受物件轉向及模糊等物件變形影響。然而因深度學習方法需要訓練神經網路中大量的方程式權重，因此相較傳統機器學習方法，需要更多的訓練資料及計算時間。目前物件偵測大量使用卷積神經網路（convolutional neural network, CNN），為當今

影像處理以及深度神經網路的主流發展技術。

目前的物件偵測技術能夠利用單一模型偵測上千種不同種類的物件。人臉辨識、行人偵測均是此技術的主流應用方向；而後續將介紹的物件追蹤技術也是物件偵測技術的延伸應用課題。以下將介紹物件偵測之模式。

YOLO9000: Better, Faster, Stronger

YOLO9000 [1] 係由同團隊所發表的前版本 YOLO [2] 改進而來，在 2017 年時發表於 CVPR 研討會的物件偵測技術。

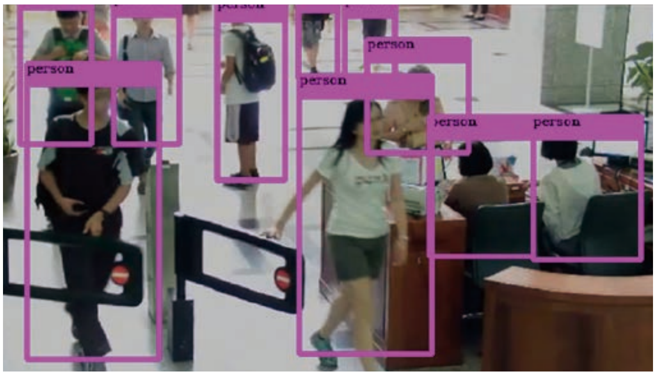
YOLO 整個模型僅使用一個 CNN 網路來產生定界框（bounding box）與進行物件分類；因其模型構造簡單，係屬於端對端（end-to-end）的技術方法，所以比較容易進行訓練，辨識速度也快。YOLO9000 較先前的版本運算時間更為縮短，且精準度提升，同時經過訓練可以辨識高達 9000 種的不同的物件。因在較高的精準度之下同時又能有很快的運算速度，例如適應較差解析度的版本 YOLOv2 288 × 288 之運算速度可高達每秒 90 幀，因此 YOLO9000 非常適合應用於即時影像辨識。

圖 1 為本研究測試 YOLO9000 之辨識成果，(a) 至 (c) 之底圖分別為國立臺灣大學圖書館監視影像、國立臺灣大學土木系女籃合影及公館捷運站以手機側錄之影像，(d) 則係由網路上採集因災所破壞之房屋影像。圖 1 中 (a) 及 (b) 表現 YOLO9000 有能力偵測出不同角度的人；(c) 中呈現此演算法可偵測出重疊的物件，如圖中的人及自行車；而 (d) 則呈現 YOLO9000 能夠辨識建物被破壞的狀態。

Mask R-CNN

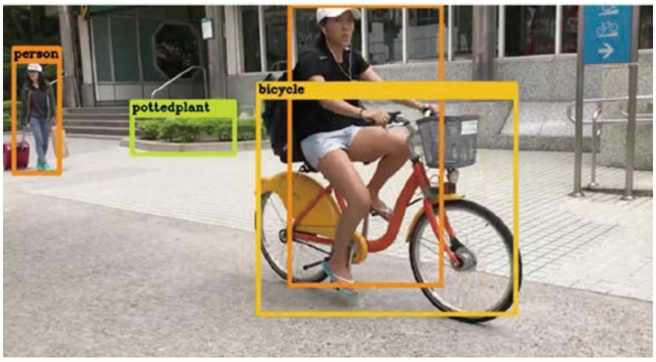
Mask R-CNN [3] 係於 2017 年發表於 ICCV 研討會的影像辨識技術。有別於以往個偵測技術，只會產出被偵測物體的定界框，Mask R-CNN 會進一步產出對應物體形狀的精準遮罩（mask），同時還可對人物產出身體結構。本研究將 Mask R-CNN 應用於臺北市捷運站月台內側錄之影像，如圖 2 所示，顯示 Mask R-CNN 優良之人體姿勢估計技術。

Mask R-CNN 演算法是以 Faster R-CNN [4] 為基礎進行改進的版本，其在原始的 Faster R-CNN 結構上增加了一產生遮罩的分支，此分支與產生定界框的神經網路平行。而為了產出精準的遮罩，將 Faster R-CNN 中負責產生可能含有物件之後選區域的 RoIPooling 替換為 RoIAlign，以解決 RoIPooling 在執行 Max Pooling 時造成的影像偏移現象。



(a) 臺大圖書館監視影像

(b) 臺大土木系女籃合影



(c) 公館捷運站

(d) 因災所破壞之房屋

圖 1　YOLO9000 辨識成果

圖 2 Mask R-CNN 辨識成果 — 臺北市捷運站月台內側錄之影像

物件追蹤

物件追蹤係指對影片中單一或多個物件進行追蹤的技術，被追蹤之物件可以是人、動物、車輛，也可以是人的部分特徵，如頭部、四肢關節等。而目前主流的追蹤技術通常包含兩個步驟：一是物件偵測；二是對將影片前後影格的所偵測到的物件進行配對。而配對的目的，是希望能夠串聯同一個物體於影片中不同的影像位置，進而得到相同物件之軌跡或進行追蹤。而前述章節所介紹的 YOLO 以及 Mask R-CNN 均是追蹤技術中常利用的物件偵測工具。

Simple Online and Realtime Tracking

Simple Online and Realtime Tracking [5]，後稱 SORT，為一能處理線上即時追蹤課題的人物追蹤技術，於 2016 年發表於 ICIP 研討會。

本演算法利用 Faster Region CNN（FrRCNN）偵測影片中各幀行人的定界框，以卡門濾波（Kalman filter）與行人隨著時間的移動資料，預測行人於最新影格的定界框。在獲得預測所得的定界框與最新影格所偵測到的定界框後，利用匈牙利演算法（Hungarian algorithm）比較所有組合之預測定界框與偵測定界框倆倆之重疊比率(intersection-over-union, IOU)，將前後影格所偵測到的人物進行配對。

圖 3 為本研究於 2012 年跨 2013 年時於臺北市政府捷運站出入口側錄大量行人進入捷運閘口之影像，並利用 Mask R-CNN 抓取人體姿勢後，以 SORT 進行配對與追蹤之成果。圖 3 中之偵測結果，為作者以 SORT 為架構，將 FrRCNN 替換為 Mask R-CNN 之成果。

PoseTrack: Joint Multi-Person Pose Estimation and Tracking

PoseTrack: Joint Multi-Person Pose Estimation and Tracking [6] 為一行人追蹤技術，能夠處理多人身體姿勢偵測以及追蹤，如圖 4 所示，其中，圖 4 底圖亦來自臺北市捷運站月台內側錄之影像；此技術於 2017 年發表於 CVPR 研討會。

PoseTrack 先利用神經網路 RastNet-101 偵測可能的關節點，包含人的頭部、四肢以及重要關節之位置；再以 Non-maximum Suppression 以及 Confidence of Spatial Edge Candidates 判斷出最佳的關節位置。在獲得關節位置後，PoseTrack 利用數學規劃中之線性整數規劃法，以前後影格特徵點（頭部與關節）之相連路徑最短為目標值，進行最佳化之求解。

此演算法因使用線性整數規劃求解，因此相較其他方法速度較為緩慢，無法即時完成運算，同時在求解上也有人物數量限制。然而偵測結果精準，且能克服人物重疊或暫時被遮蔽問題之優點。

Detect-and-Track: Efficient Pose Estimation in Videos

Detect-and-Track: Efficient Pose Estimation in Videos [7]，為一人物追蹤技術。本文將以 Detect-and-Track 作為其簡稱。此方法能夠處理多人身體姿勢偵測以及追蹤，於 2018 年發表於 CVPR 研討會。

此演算法物體偵測部分修改了 Mask R-CNN，將其增加一時間維度，以同時輸入固定時間區間內的多幀影格作為偵測參考，稱為 3D Mask RCNN。在獲得偵測

圖 3 SORT 應用於臺北市政府捷運站出入口側錄追蹤之成果

圖 4　PoseTrack 於臺北市捷運站月台內側錄影像辨識與追蹤之成果

成果後，此演算法比較多種配對方法，其中包含比對人物所提取特徵的相似程度、利用 IOU 比對人物位置的相似程度，以及比對人物身體姿勢的相似程度等。最後獲得比對人物特徵所得之配對效果最好。

圖 5 之底圖來源為 MOTChallenge（Multiple Object Tracking Challenge），名稱分別為 TUD-Campus 與 MOT17-09 的原始影像，本研究應用 Detect-and-Track 追蹤影像中之行人，如圖所示，被此演算法判定相同的人物會以相同顏色的外框顯示追蹤結果。

小結

物件追蹤與偵測是兩個有關聯但不同的影像處理技術：物件偵測只負責發現影像中的特定物體；而追蹤則是找出偵測物體在影片中的動向。隨著深度學習技術以及運算硬體的發展，物件追蹤與偵測技術都越來越精確，運算時間縮短，能夠處理的物件種類隨之增加。

追蹤與偵測技術在土木水利方面預期將可有不少應用空間，例如影像偵測技術可應用於地震等大型災害發生後受損建物偵測；若結合無人飛行載具獲得高空拍攝影像，則可幫助相關單位在短時間內判斷一地的建物受損、邊坡改變甚至河水暴漲等概況。而追蹤技術可應用於車流、人流甚至河流等流向資訊之取得，減少相關資料的搜集成本。

即時之同步定位與地圖構建

本章節將先介紹即時之同步定位與地圖構建技術之最新研究成果，再呈現影像深度估計之研究。進而討論在土木及水利工程之可能應用。

即時之同步定位與地圖構建

即時之同步定位與地圖構建技術指利用一移動偵測器在一未知的空間中進行移動偵測，透過偵測器獲得的資料（例如影像資訊）進行分析，構建地圖並定位偵測器之位置。現今之同步定位與地圖構建（Simultaneous Localization and Mapping, SLAM）技術普遍追求能即時產生地圖構建與定位結果。

ORB-SLAM2: an Open-Source SLAM System for Monocular, Stereo and RGB-D Cameras

ORB-SLAM2 [8] 係一 SLAM 技術，其改良自前版本 ORB-SLAM [9]，能夠支援一般單鏡頭影像、立體攝影機以及彩色深度攝影機的影像分析；並且能夠在多種不同的環境執行地圖構建，包含室內以及室外的空

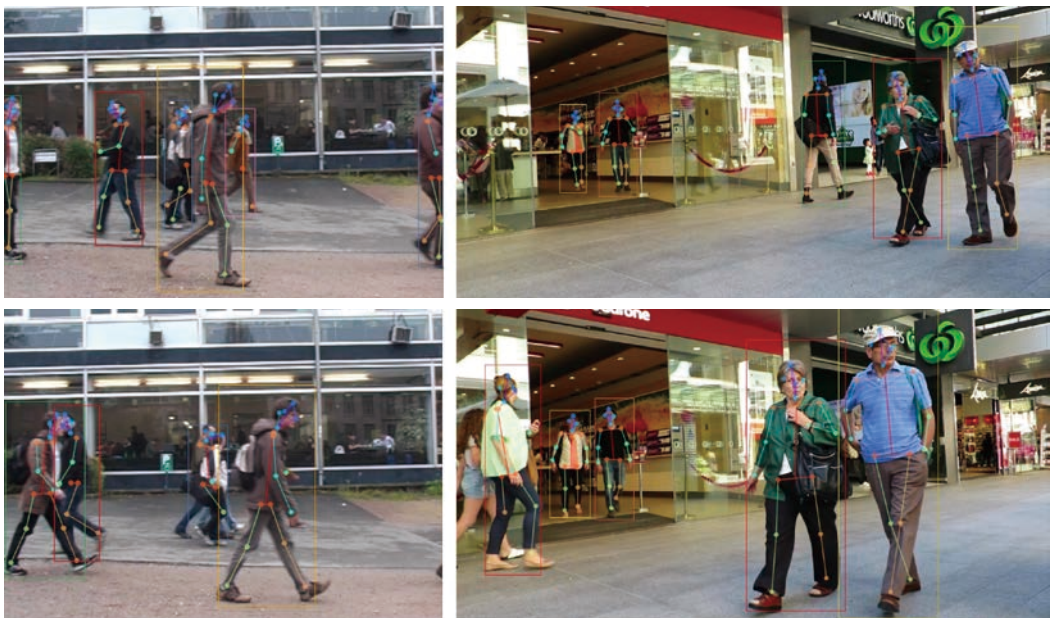


圖 5　Detect-and-Track 辨識與追蹤成果

拍場景，可以與手持攝影機結合，另外也可以搭載於無人飛行器上進行大範圍的地圖構建。由於其不耗費大量運算資源，因此可以運行於 CPU 環境，同時也可支援即時的分析。

圖 6 為 ORB-SLAM2 的執行成果，其底圖為本研究於國立臺灣大學土木系 3 樓走廊側錄之影像。左側為輸入之偵測影像，其中綠色標註點為影像中偵測到的特徵點。而右側則為 SLAM 輸出之建模成果，其中紅色點為特徵點對應至 3D 空間點雲，藍色方框則為偵測器（此例為手持攝影機）之定位呈現。

深度估計

影像的深度估計（Depth Estimation）技術是指利用影像估計影像中各物體的深度，過往的技術主要利用雙鏡頭、多鏡頭，在得知鏡頭相對位置的情況下，透過數學模型的方式求得。然而隨著深度學習技術的發展，越來越多研究團隊投入單鏡頭影像深度估計的研究領域。

Unsupervised Monocular Depth Estimation with Left-Right Consistency

左右一致之非監督式單一鏡頭深度估計（Unsupervised Monocular Depth Estimation with Left-Right Consistency）[10]，後稱 monoDepth，為一影像深度分析技術，可以利用單一鏡頭的影像，透過訓練過的神經網路，進行影像中的各部分進行深度的推估，成果如圖 7 所示。該研究發表於 2018 年 CVPR 研討會。

一般的深度分析技術多半為監督式的，缺點為需要大量有深度標籤的影像資料，有極高的資料搜集成本。該研究利用雙鏡頭的攝影機，經過運算獲得大量的深度資料，透過此資料訓練出評估單鏡頭影像深度的神經網路，解決需要大量標籤資料的成本問題。

此技術預計可以與 SLAM 技術結合，使得可以在僅有單一鏡頭的載具也可以獲得合理的影像深度資訊，增加構建地圖的精準度。

小結

SLAM 技術使得一偵查器可以自動於未知空間內進行定位與環境地圖構建，而隨著深度學習技術以及運算硬體的發展，偵查器得以搭載更精確、更快速的演算法。

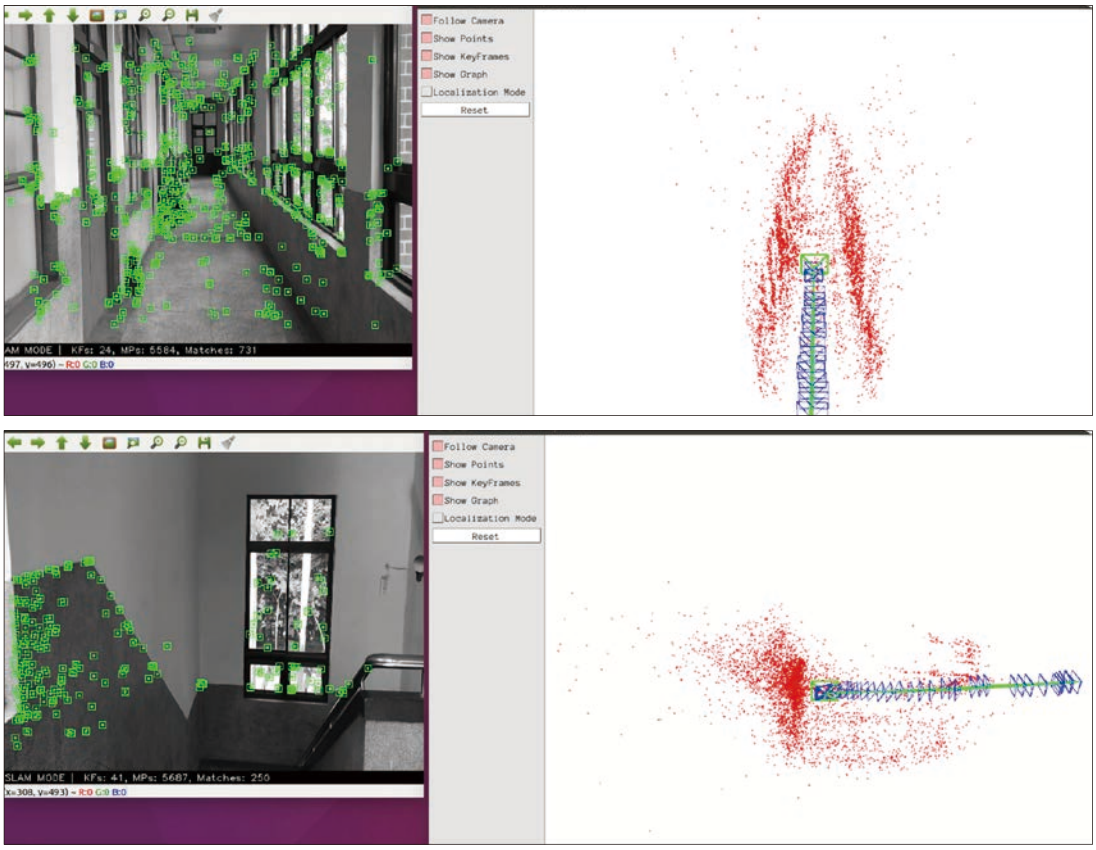


圖 6 ORB-SLAM2 定位與空間建構於臺大土木系系館三樓之展示成果

SLAM 技術預期可在土木水利以及防災領域有廣泛的應用。除了可應用於未知建物及環境的建模外，也可藉由結合 SLAM 技術的偵查機器人，執行倒塌建物內的情資搜查與建物建模等任務，此偵查機器人可結合簡單醫療或物資之投放功能，以解決倒塌建物中通道狹窄之問題，保障救災人員之自身安全，並完成物資、醫療資源之快速配送。

影像風格轉換

此章節將針對影像風格轉換作一介紹及說明。生成對抗網路（Generative Adversarial Network, GAN）是一種非監督式學習的神經網路演算法。生成對抗網路的結構包含兩組神經網路，一組為產生器，負責產生所需的資料，一組為辨別器，負責辨別產生器所產生的資料為真實資料或是產生器的假資料；其結構如圖 8 所示。透過反覆訓練產生器以及辨識器，使得產生器有能力輸出與目標需求相當接近的影像。

利用生成對抗網路，可將一筆資料（例如影像、影片、聲音檔等）輸入轉換成目標需求之風格。

Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks

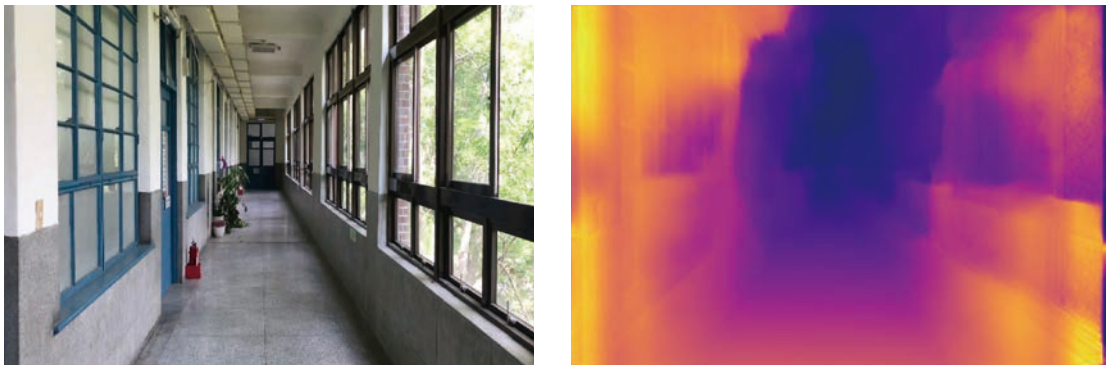
Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks [11]，後稱 Cycle GAN，為一生成對抗神經網路，此演算法之設計使其可將一輸入影像風格轉換產生一輸出影像，例如將一照片轉換成梵谷繪畫風格的影像。此演算法於 2017 年發表於 ICCV 研討會。

Cycle GAN 將基本的生成對抗神經網路為架構，為了避免對抗生成網路的產生器無視輸入影響逕行輸出與輸入影像無關的照片，在演算法中引入另一組產生器，負責將輸出影像再次轉換成原始輸出影像，並要求輸入影像必須要非常接近還原影像，迫使影像風格轉換的產生器需保留一定程度的原始影像要素。

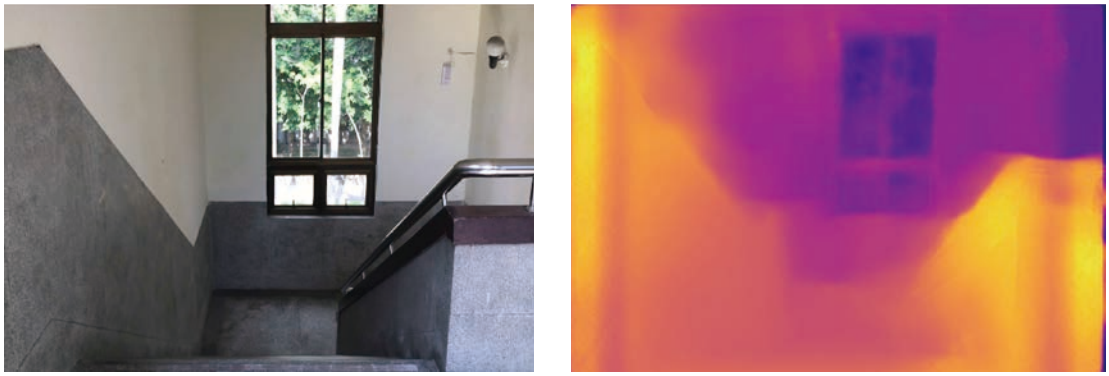
圖 9 為本研究利用 Cycle GAN 輸出之成果，底圖為本研究於國立臺灣大學圖書館正門及第一學生活動中心正門所拍攝之影像。如圖所示，透過 Cycle GAN，可將一照片變換其色澤，或將照片轉換成油畫風格之影像。

Image-to-Image Translation with Conditional Adversarial Networks

Image-to-Image Translation with Conditional Adversarial Networks [12]，後稱 Pix2pix，為一生成對抗神經網路，此演算法之設計使其可將一輸入影像風格



(a-1) 室內空間影像 a　(a-2) 室內空間影像 a 之深度分析結果



(b-1) 室內空間影像 b　(b-2) 室內空間影像 b 之深度分析結果

圖 7　monoDepth 影像深度分析於臺大土木系系館三樓之展示成果

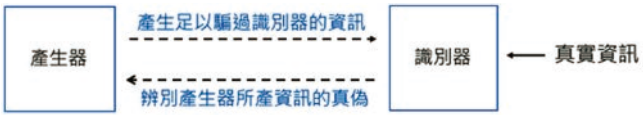


圖 8 生成對抗網路基本結構示意

轉換產生一輸出影像，例如將空照圖轉換成地圖，或將黑白照片轉換成彩色照片。此演算法於 2017 年發表於 CVPR 研討會。

本研究團隊於 2017 年發表 Using Context Encoders in AEC/FM [13] 於 ICCCBE 研討會。該研究將 cGAN 演算法之概念應用於影像補遺，消除在照片中的行人影像，同時根據消除區域周圍的資訊預測遺失的影像。圖 10 為本研究將 cGAN 應用於影像補遺之成果，其底圖為本研究於國立臺灣大學圖書館正門拍攝影像之成果。

小結

生成對抗網路為近期最受矚目的演算法之一，有別於以損失函數（Loss Function）為目標進行訓練，生成對抗網路能夠衡量圖像整體結構，進而產生以假亂真的輸出成果。例如 NVIDIA 團隊利用生成對抗網路產生的超高解析度虛擬人像 [14]。

在土木及水利工程的應用上，可利用影像風格轉換技術，將視線不佳、模糊之影像，轉換成其清晰的版本，以作為資訊搜集、影像判讀之輔助。另外也可利用其將被擋住物件進行補遺。而隨著人工智慧技術之精進，未來可期待利用此技術將有遮蔽之建築外觀輸入，生成其對應之主要設施圖，可用於土木與水利關鍵營建專案之規劃、設計、建造及營運維護等階段。

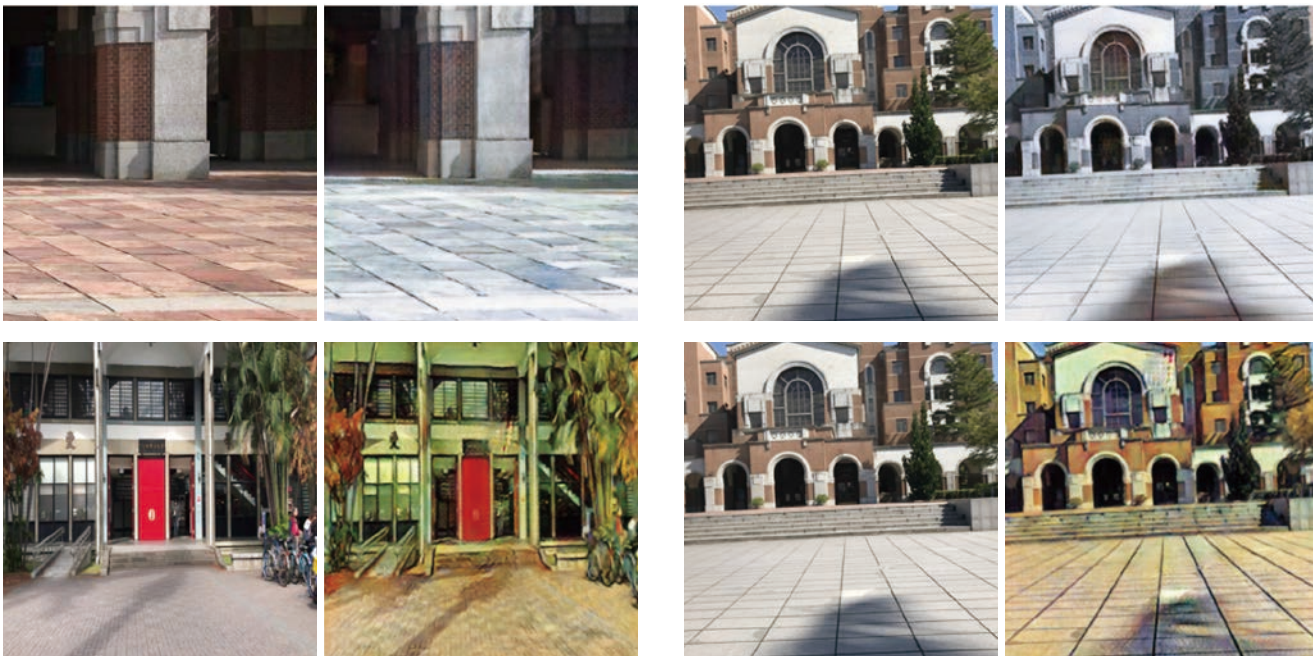


圖 9 於臺大校園之原始輸入影像與 Cycle GAN 之輸出成果

圖 10 cGAN 應用於臺大校園之影像補遺

結論

本研究由 2017 年至 2018 年之 CVPR 研討會文章中彙整出對於建築、工程與營建及設施管理領域深具潛在貢獻之最新技術，包含針對近期於物件偵測、物件追蹤、即時之同步定位與地圖構建、深度估計與影像風格轉換。

對於上述技術，本團隊已利用臺灣大學所拍攝之影像，進行初步之研究及應用，並將影像風格轉換中 image-to-image 技術實際應用於影像補遺中。另分別對物件偵測、物件追蹤、即時定位與地圖構建與深度估計提出應用之建議，相信對土木及水利工程應深具其應用價值。

參考文獻

1. J. Redmon and A. Farhadi (2016), "YOLO9000: Better, Faster, Stronger", 2016.
2. J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "(2016 YOLO) You Only Look Once: Unified, Real-Time Object Detection", Cvpr 2016.
3. K. He, G. Gkioxari, P. Dollar, and R. Girshick (2017), "Mask R-CNN", In Proceedings of the IEEE International Conference on Computer Vision.
4. S. Ren, K. He, R. Girshick, and J. Sun (2015), "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks" , pp. 91–99,.
5. A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft (2016), "Simple online and realtime tracking", Proc. - Int. Conf. Image Process. ICIP, vol. 2016–Augus, pp. 3464–3468.
6. U. Iqbal, A. Milan, and J. Gall (2017), "PoseTrack: Joint Multi-Person Pose Estimation and Tracking", In IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
7. R. Girdhar, G. Gkioxari, L. Torresani, M. Paluri, and D. Tran (2018), "Detect-and-Track: Efficient Pose Estimation in Videos", In CVPR.
8. R. Mur-Artal and J. D. Tardos (2017), "ORB-SLAM2: An Open-Source SLAM System for Monocular, Stereo, and RGB-D Cameras", IEEE Trans. Robot.
9. M. J. M. M. Mur-Artal Raúl and J. D. Tardós (2015), "{ORB-SLAM}: a Versatile and Accurate Monocular {SLAM} System", IEEE Trans. Robot., vol. 31, no. 5, pp. 1147–1163.
10. C. Godard, O. Mac Aodha, and G. J. Brostow (2017), "Unsupervised Monocular Depth Estimation with Left-Right Consistency", In CVPR.
11. J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros (2017), "Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networkss", In Computer Vision (ICCV), 2017 IEEE International Conference on.
12. P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros (2017), "Image-to-Image Translation with Conditional Adversarial Networks", CVPR.
13. S.-Y. Wen, A. Y. T. U. Chen, Y.-F. National, and T. U. Chiu (2018), "Using Context Encoders in AEC/FM".
14. T. Karras, T. Aila, S. Laine, and J. Lehtinen, "PROGRESSIVE GROWING OF GANS FOR IMPROVED QUALITY, STABILITY, AND VARIATION".