◎ 陳冠州、蔡家齊、曾慶鎧、林冠廷、郭峻因

AI 自駕車的物件偵測

近來發展的人工智慧技術不但有偵測物件的能力，
還會分析並判斷物件的行為模式，
使得無人操作的自動駕駛車指日可待。
這種物件偵測、追蹤及判斷的能力究竟是如何達到的？

智慧型車輛發展的目的，
是為了達到所有汽車製造廠的共同目標：安全、便利、舒適、環保。

隨著半導體及電子技術的發展，汽車製造商及零組件系統廠商巧妙地把電子技術應用在汽車設計及製造上，使得汽車從過去的封閉系統轉變成能與外界溝通的智慧型車輛。而智慧型車輛發展的目的，是為了達到所有汽車製造廠的共同目標：安全、便利、舒適、環保。

《*EE Times*》於 2013 年下半年多次報導車用電子系統技術發展的趨勢，包含預測具有全部或部分自動駕駛功能的汽車將在 2025 ～ 2030 年達到 25 ～ 30%的市場占有率、介紹未來汽車電子系統所採用的手勢操控技術等。由於 Google 自動駕駛車輛成功挑戰歐亞長距離行駛，加上 BOSCH 研發自動駕駛汽車技術的成果，使得世界各國車廠對於自動駕駛汽車技術更具信心，也帶動了各式輔助汽車駕駛人的先進自動駕駛技術的蓬勃發展。

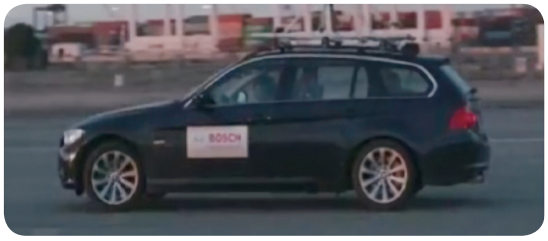
自動駕駛的關鍵技術之一的人工智慧技術，近幾年來也愈趨成熟，本文介紹目前應用於駕駛輔助與自動駕駛的主流深度學習演算法與實際成果。

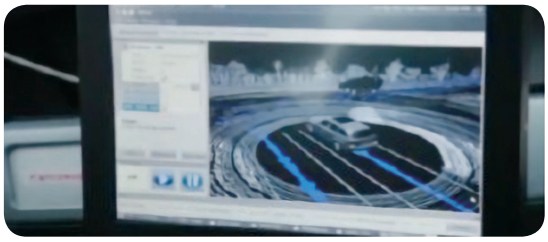
深度學習

卷積類神經網路　在深度學習的架構中，卷積類神經網路（convolutional neural network, CNN）是相當受歡迎的一個架構。1989 年由 LeCun 等人提出的 CNN 架構，在手寫辨識分類或人臉辨識方面都有不錯的準確度。

近年來，隨著 CPU 效能的提升與繪圖晶片平行化技術的發展，讓具高複雜度、費時的深度學習演算法在即時應用上露出曙光，透過繪圖晶片可讓訓練模組與測試的時間大幅縮短。伴隨著得以取得多樣的影像資料庫，CNN 可觸及更多在照片與影片

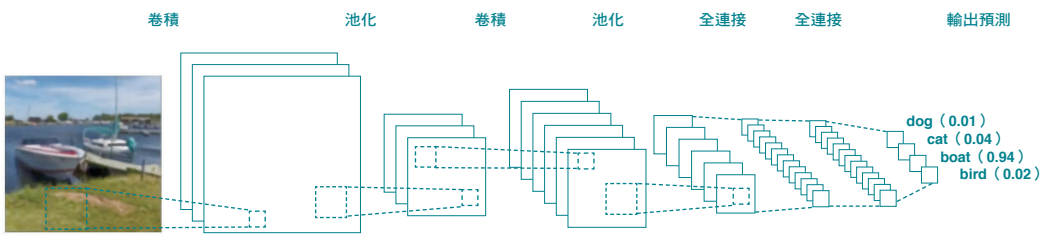
Google 自動駕駛智慧車與 BOSCH 自動駕駛汽車雛型





深度學習常用名詞	解釋
Convolution layer	可以提取影像的特徵，使深度學習模型理解。
Pooling layer	把輸入資料的空間維度下降以降低運算複雜度，加速深度學習模型運算。
Fully Connected layer	最後深度學習模型用於判斷物件種類、位置等。

Convolutional neural network（CNN）的基本架構



上的應用。例如近來接續發表的 AlexNet、ZF-Net、VGG Net、GoogLeNet 等，在精確度與效能上都有所改善，甚至在有些情況中可以超越人眼可辨識的範圍。

在影像上的技術發展　深度學習在影像應用上正蓬勃發展，從物件分類、物件偵測、物件追蹤、行為分析至反應決策，無一不朝向提高準確度和效能的方向發展。以下介紹近年來在處理物件分類與物件辨識方向熱門的 CNN 網路架構與改進。

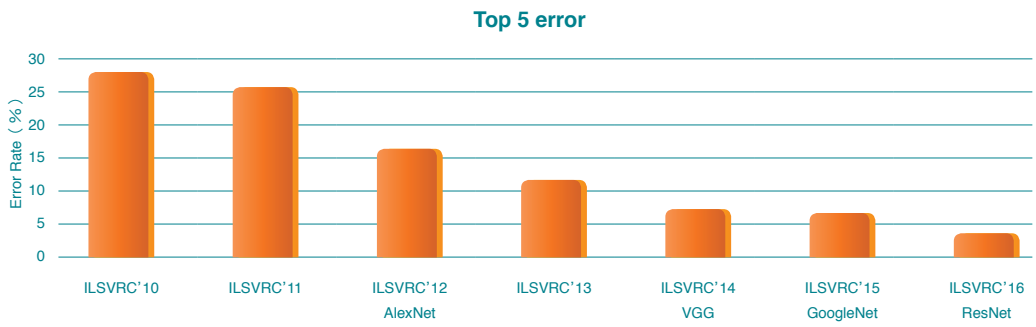
物件分類　物件分類是分析一張照片中包含的物件種類，主要是先使用 convolutional layer 進行特徵擷取，再經由 fully-connected layer 合併特徵進行判斷。而在深度學習網路優劣評比中，ILSVRC（ImageNet Large Scale Visual Recognition Competition）是一種標竿排名比賽，方便研究者評估與比較物件偵測以及影像分類演算法。以下是幾個著名影像物件分類的網路架構：

LeNet —這是首先成功的 CNN 架構，由 LeCun 在 1990 年提出，見長於辨識數字和英文字母。

AlexNet —第一個讓 CNN 網路架構開始在電腦視覺中蓬勃發展的網路，由 Alex Krizhevsky、Ilya Sutskever 和 Geoff Hinton 提出，並在 2012 年的 ILSVRC 比賽中比第二名取得了大幅度的領先（Top 5 error 16%，第二名是 26%）。AlexNet 的網路架構類似於 LeNet，但更深、更大，並且開始使用多個層疊的 convolutional layer，然後再連接 pooling layer，有別於以往一層

深度學習在影像應用上正蓬勃發展，從物件分類、物件偵測、物件追蹤、行為分析至反應決策，無一不朝向提高準確度和效能的方向發展。

歷屆 ILSVRC 冠軍 Top 5 的 Error Rate



convolutional layer 都會馬上連接一層 pooling layer 的架構。

ZF-Net —由 Matthew Zeiler 和 Rob Fergus 所提出，並在 2013 年的 ILSVRC 取得優勝。他們提出了一個把 CNN 網路中間的特徵層取出並視覺化的方法，便於分析 CNN 架構不足的地方並加以改進。ZF-Net 便是基於 AlexNet 的優化，活化 AlexNet 中無用的特徵，以得到更好的特徵擷取和辨識效果。

VGGNet —由 Karen Simonyan 和 Andrew Zisserman 提出，最主要的貢獻是證明了 CNN 網路的深度對準確度的影響，愈深的網路提供愈好的準確度。但 VGGNet 的網路架構需要更高的計算複雜度，以及更高的記憶體需求。

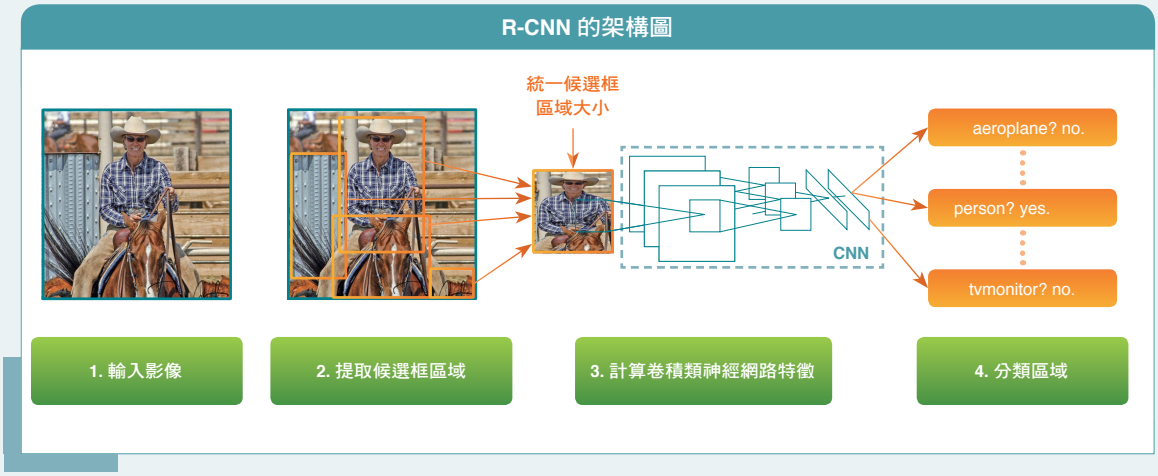
GoogLeNet —由 Google 提出，在 2014 年的 ILSVRC 中取得優勝。GoogLeNet 提出了 inception module，可以同時結合不同 level 的特徵，並可串連不同 scale 下的特徵提取值，讓網路可以更深，同時減少參數（例如 GoogLeNet 使用 400 萬個參數，AlexNet 使用了 6,000 萬個參數，而 VGGNet 更需要 14,000 萬個參數），並擁有更好的辨識效果。

ResNet— 由 Kaiming He 等人提出，在 2015 年的 ILSVRC 中得到優勝。ResNet 提出的架構可讓特徵值有捷徑跳至後幾層，讓 CNN 網路得以更深，並大量使用 batch normalization，是目前最佳技術的 CNN 分類網路架構，但它的計算複雜度也最高。

物件偵測　相較於物件分類，物件偵測的挑戰更加艱難，它是在一張影像中，需要同時定位物件的座標，再做出分類。現有技術已從最早期開始的遍數法，也就是把影像中所有可能性都使用 CNN 網路判斷，到後來提出在辨識上能更有效率的物件找尋方式。在評估物件偵測演算法優劣上，一般使用 PASCAL VOC （visual object classes）這個開源的標準資料庫進行測試。以下介紹目前著名的物件偵測技術。

Sliding windows —早期較原始的找尋目標方式，先把一張圖片由小到大的視窗，整張影像全部掃過一遍，並擷取掃過的影像，餵進 CNN 網路分類。這個方法簡單，但計算量非常大，不適合即時應用。

Region proposal CNN network —這個技術是先對影像進行區域提取，透過演算法把影像切分為可能含有物件的區域，再擷取這些區域，提供給物件分類的 CNN 網路判別。



著名的架構有 RCNN、Fast-RCNN、Faster-RCNN 等。RCNN 使用 selective search 演算法進行區域提取，經過演算後，可以把一張影像取出許多個有可能的區域。但這演算法過於複雜，並且每個區域中進行 CNN 特徵提取時會重複計算，進而導致效能瓶頸。

Fast-RCNN 一改良自 RCNN，加速了 RoI pooling layer，使得 RoI pooling layer 可以把不同大小的輸入 mapping 到固定大小的層，並且改動 RCNN 的流程，先提取可能區域，然後做特徵提取，再從提取完成的特徵圖進行分類。這技術可以避免 RCNN 中特徵提取重複計算的問題，在保證精確度下提升運算速度。

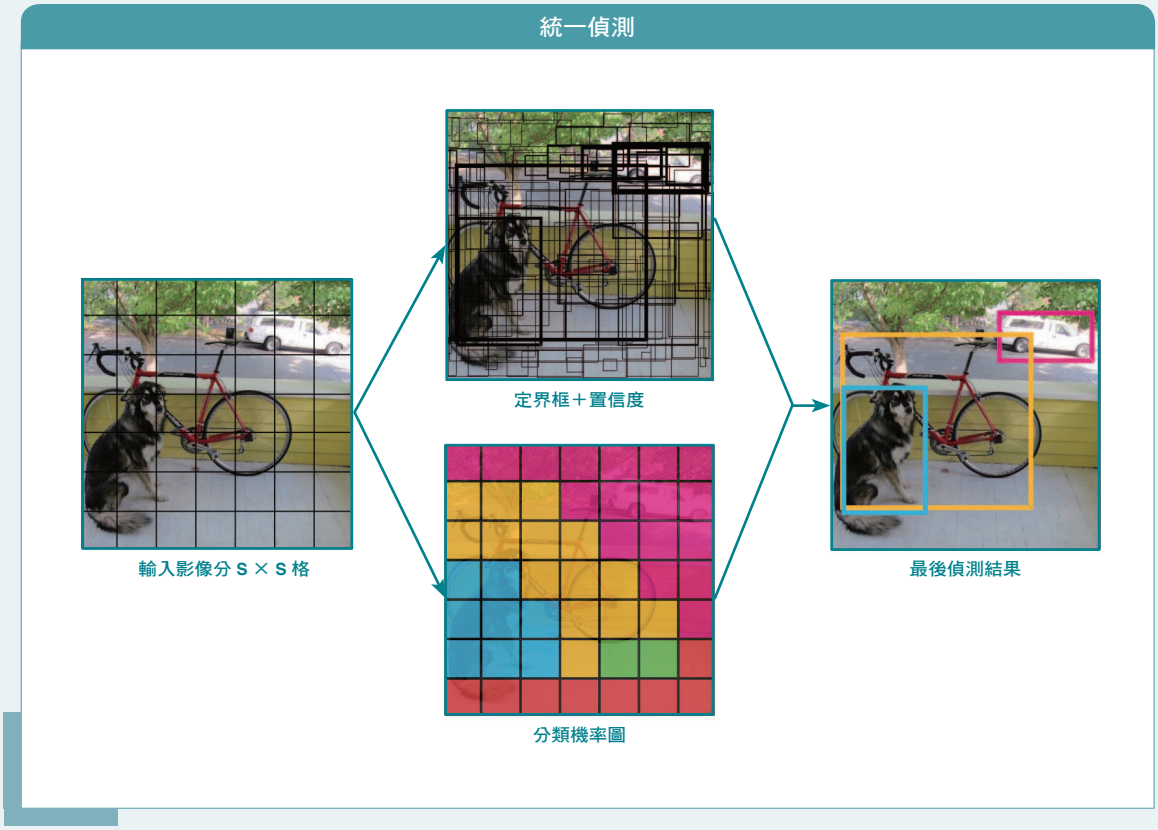
Faster-RCNN — Faster-RCNN 更進一步使用 region proposal network（RPN）取代原先的 selective search，把可能區域的提取方式內嵌到 CNN 網路中，提供訓練和測試一個 end-to-end 的網路架構。RPN layer 也改善了原先 selective search 只使用 CPU 運算的問題，把可能區域提取透過繪圖晶片加速。

在 regional proposal CNN network 方面，目前常見的著名網路架構有 ZF+Faster-RCNN、VGG16+Faster-RCNN、R-ResNet-101+Faster-RCNN、PVANet 等，前兩個是使用 ZFNet 和 VGGNet 再加上 Faster-RCNN 架構產生的物件偵測網路。

ResNet-101+Faster-RCNN 是目前準確度最高的架構，準確率達到 83.8％，但整體運算量非常大。PVANet 則多導入了 C.ReLU module 以及改進了 inception module，透過分析特徵層的特性，讓計算複雜度下降的同時擁有更好的精準度。在 PASCAL VOC2012 中有 82.5%的精準度，但運算量只有 ResNet-101+Faster-RCNN 的約十分之一。

統一偵測—不同於 region proposal CNN network，統一偵測（unified detection）對於物件偵測的方式不先提出可能區域再進行分類，而是直接把可能區域提取的方式轉為回歸問題。它透過預先設定好的幾個 bounding box，利用 CNN 網路進行 bounding box 位置回歸以及可信度判斷，同時進行分類。這方法可大幅提升物件偵測的速度，但對於小的物件以及準確度仍有待改進。以下介紹幾個著名的 unified detection 物件偵測網路：

You only look once（YOLO）— 如其名，人眼在分別物體時並非先抓取位置再進行判斷，而是看到物體的同時辨識物件。YOLO 提出了 unified detection 的物件偵測方式，透過預先設定好的 bounding box，再透過縮放平移去貼近到物件邊緣同時判斷，因而大幅提升速度。但它的準確度尤其是對於較小的物件，表現較差。



Single shot multibox detector 一改良自 YOLO 網路架構，它把網路分為兩個結構：feature extraction 和 auxiliary。Feature extraction 的部分與一般網路類似，用於特徵提取，auxiliary 則是把提取出的特徵再進一步降低維度，讓最後的 fully-connected layer 同時結合不同維度的特徵，進行 bounding box 回歸和物件分類。相較於 YOLO 只使用單一維度的特徵進行判斷，這種方法可以有更佳的準確度，在 PASCAL VOC 有 82.2%的平均準確度。

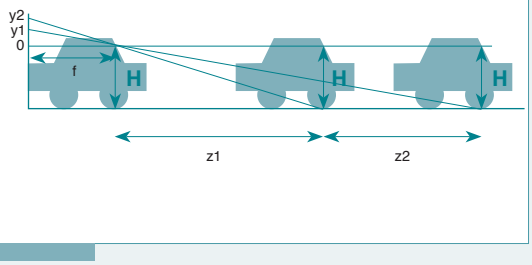
基於物件偵測的自駕車應用

深度學習演算法擁有良好的精準度和穩定性，但伴隨的是較高的計算複雜度。然而這個演算法可以大量地平行化，因此適合利用繪圖晶片加速演算。訓練的策略也是機器學習很重要的一環，且要能夠在嵌入式系統上實現，因此它的網路架構必須設法精簡。自駕車在路上容易遇到的物件有車輛、機車騎士、行人等，鎖定這幾類物件偵測可以降低深度學習的複雜度，使得在同樣的精準度下達到更快的偵測速度。

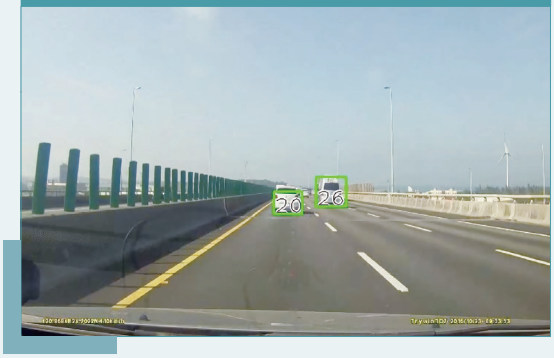
Pascal VOC2007 datasets 中包含 20 類物件，如飛機、腳踏車、鳥、船、貓、狗等，自駕車所需的目標只需要偵測汽車、行人與機車騎士。由於機車騎士具備行人特徵，可由行人樣本來偵測，因此只需要從 Pascal VOC2007 datasets 的 20 類樣本中取其中兩類，汽車與行人，當作自駕車系統的一部分訓練樣本，就可訓練出自駕車所需要的模型。

深度學習演算法擁有良好的精準度和穩定性，但伴隨的是較高的計算複雜度。

車距估測方法的示意圖



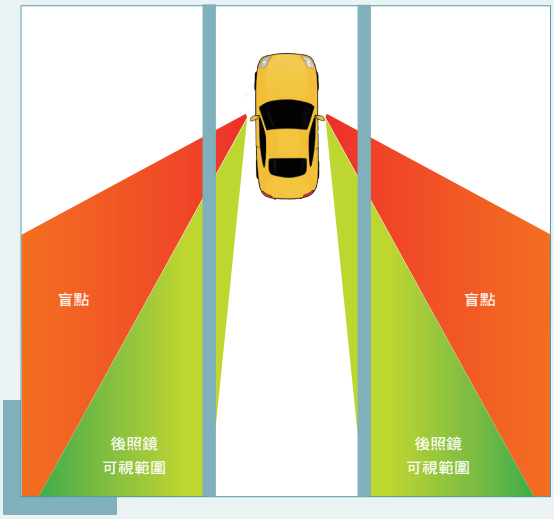
前車防碰撞系統的執行結果



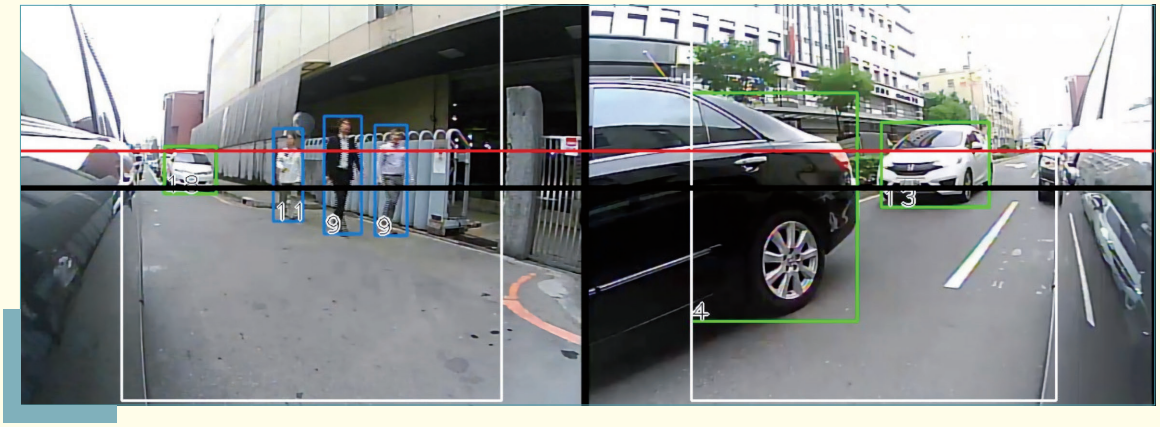
距離估測的結果範例

距離偵測與前車防撞警示　車距可以利用鏡頭水平拍攝後，藉由鏡頭的高度與焦距推算。在前車防撞警示方面，可利用車前方的相機擷取影像後，透過深度學習物件偵測演算法偵測物件。再由前述車距估算的方法對物件位置分類以及距離估算，並根據自駕車與前方車輛的距離調整安全距離，以避免前方車輛突然緊急煞車，導致自駕車煞車不及而追撞前方車輛。

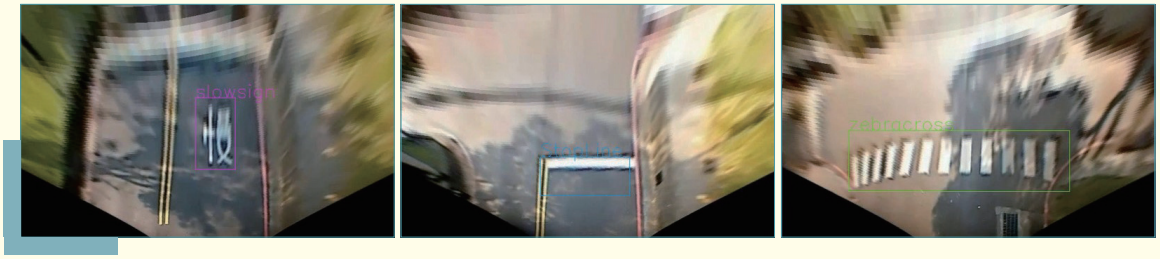
盲區危險警示　駕駛人開車時，變換車道或轉向都應注意左右方車輛。而一般車輛在後照鏡的視覺上都有盲點，唯有透過轉頭才能注意到盲點區域的車輛。但在駕駛時轉頭又容易偏離車道或無法注意前方車況，同樣地自駕車也需考慮這問題。



盲點區域示意圖，後照鏡看到的範圍是黃色區域，紅色區域則是駕駛座位的盲點。



車側盲點警示系統的基準線（紅色線）與警示結果。



路面標線標字偵測結果，左圖識別出標字「慢」，中圖識別出停止標線，右圖則識別出斑馬線。

解決方案是藉由 AI 深度學習偵測物件，準確地辨識出左右後方區域中的物件，再整合所有資訊，透過鏡頭預先設定的基準線，可以判斷出偵測到的物件是否在需要警示的位置，並以相較於自身車輛的距離而警示。

應用於路面標線標字偵測　由於許多事故都是因汽車駕駛未遵循路上的標線或標字行駛而造成，因此當自駕車行駛在路上時須偵測並理解標線及標字。為使深度學習演算法訓練與偵測更加穩健，通常把路面由俯視轉為鳥瞰角度，建構可應用於馬路標線和標字偵測的模型，讓自駕車遵行路上的標線標字內容，而能安全地行駛在道路上。

Level-5 自駕車

隨著深度學習演算法不斷的進步，從 2012 年 AlexNet 贏得 ILSVRC-2012 冠軍，到 2016 AlphaGO 以四比一的成績贏南韓棋王李世乭，人工智慧已成為 21 世紀最夯的顯學。國際大廠都以 Level-5 自駕車為目標開發，使得深度學習演算法有十足能量繼續突破。在可見的未來，深度學習不只有偵測物件的能力，還會分析並判斷物件的行為模式，無人操作的自動駕駛車指日可待。

陳冠州、蔡家齊
曾慶鎧、林冠廷、郭峻因
交通大學電子研究所