Journal of Science and Engineering Technology, Vol. 17, No. 2, pp. 1- 8 (2021)

基於深度學習之車輛前方障礙物距離估測

黃登淵 [1*]　謝溢樑 [2]　王清弘 [2]

[1] 大葉大學電機工程學系

[2] 大葉大學資訊工程學系

51591 彰化縣大村鄉學府路 168 號

[*]kevin@mail.dyu.edu.tw

摘　要

由於科技的進步，自主駕駛系統在可知的未來必形成一股風潮，而車輛與障礙物之間的距離估測是自主駕駛系統中一個非常重要的技術。為了達到距離估測之目的，目前發展的自主駕駛系統大都需要仰賴各式各樣的距離感測器，例如光達、雷達及超音波等，這些感測器在距離量測上通常具有高精度，但同時也伴隨著高昂價格，這將使得自主駕駛系統的推廣及普及變得愈發困難。本文提出了一個結合語義分割與深度估測之深度神經網路模型，其包含有相同卷積層數的 Encoder 與 Decoder 網路。本文所提之網路架構在 KITTI 及 Cityscapes 資料集上進行訓練，並在最後結合語義分割與深度估測等方法進行距離估測，實驗結果證實本文所提方法的可行性。

關鍵詞：人工智慧，深度估測，語義分割，深度學習

Distance Estimation of Obstacles in Front of Vehicles Based on Deep Learning

DENG-YUAN HUANG[1*], YI-JIE XIE[2] and CHING-HUNG WANG[2]

[1]Department of Electrical Engineering, Da-Yeh University

[2]Department of Computer Science & Information Engineering, Da-Yeh University

No.168 University Rd., Dacun, Changhua 51591, Taiwan, R.O.C.

[*]kevin@mail.dyu.edu.tw

ABSTRACT

Autonomous driving systems are the wave of the future; for such systems, the estimation of the distance between the vehicle and surrounding obstacles is key. Most current distance estimation methods rely on a variety of distance sensors, such as LiDAR, radar, or ultrasonic sensors. Although these sensors measure distance accurately, their high cost hinders the popularization of autonomous driving systems. To remedy this problem, this paper proposes a deep neural network (DNN) that combines semantic segmentation and depth estimation. The DNN includes an encoder and a decoder, both of which have the same number of convolutional layers. The proposed network architecture was

trained on both the KITTI and Cityscapes datasets. The proposed method provided accurate distance estimation in evaluation tests, demonstrating its feasibility.

Key Words**:** artificial intelligence, depth estimation, semantic segmentation, deep learning.

一、前言

人工智慧一直是人類嚮往的終極目標，而深度學習則是大家公認最接近人工智慧的一種技術。近年來，深度學習不管在影像辨識、語音識別，醫療診斷或自動機器翻譯等領域都有相當傑出的表現，這一切都要歸因於類神經網路的深度結構。近年來，深度學習受到大量的關注與青睞，主要來自於幾場重要的比賽，其中 2009 年採用「長短程記憶（Long short-term memory; LSTM）」 [10] 架構的深度學習網路取得手寫辨識比賽的冠軍。

電腦視覺比較常見的應用有：影像分類 [12, 29]、物體偵測 [22, 23, 27]以及語義分割 [1, 8, 9, 14]等。其中語義分割的任務是在像素等級上對整個影像進行實例分類，每個實例（或是類別）對應於影像中的物體或表示影像的一部份，例如人、車、道路及天空等。該任務也稱為密集預測（Dense prediction）。因此該任務的目標是用影像中的相應類別標記影像中的每個像素。語義分割對於場景理解非常的關鍵，可讓深度學習模型更好的學習到環境中的全域視覺背景。

對於機器人 [11]、自動駕駛 [24]、3D 環境重建 [14]及擴增實境（Augmented Reality; AR）[19]等，深度感測是必要的技術。在機器人研究領域中，影像深度是執行探索任務的重要關鍵。傳統上，有關於道路前方障礙物的偵測與距離的判斷，為了達到更可靠的感知能力，除了攝影機外，還需仰賴大量的感測元件，其中包含超音波（Ultrasound）、雷達（Radar）及 LiDAR（Light Detection And Ranging）等。本文認為在這些感測器中，基於視覺感知（Visual perception）的攝影機可提供車輛周遭環境最豐富的資訊，其中包含顏色、紋理、物體形狀以及外觀等，這些都是其他型態的感測元件所無法提供的。基於這個原因，本文提出一種基於行車紀錄器攝影機的影像感知系統，利用攝影機所獲取的影像來進行車輛前方的障礙物偵測與距離估算。

由 Long 等人 [18]所提出的全卷積網路（Fully convolutional network; FCN）是第一個端到端（End-to-end）語義分割的網路架構。FCN 使用任何大小的影像作為輸入，並輸出具有相同大小的分割影像。Long 等人 [18]首先修改了當前流行的 CNN 架構，例如 AlexNet [12]、VGG16 [29]和 GoogLeNet [30]等。在文中，他們採用卷積層以替換所有的完全連接層（Fully connected layers），藉以產生多個特徵映射圖（Multiple feature maps），因此需要上採樣（Upsampling）來讓輸入的特徵圖產生與輸入相同大小的輸出。通常上採樣是由具有大於 1 的步伐（stride）的卷積層所組成。這種方式通常又稱為反卷積或轉置卷積（Deconvolution or Transposed convolution），因為它產生的特徵圖大小大於輸入。在 FCN 中，為了優化訓練器，文中採用逐像素交叉熵損失（Pixelwise cross entropy loss）來訓練網路。此外，他們還在網路中添加了跳躍式連接（Skip connection）的結構以產生更好的輸出結果。在文中，他們使用 ImageNet 資料集來訓練語義分割模型，在 2011 年 Pascal VOC 分類挑戰（2011 Pascal VOC classification challenge）中達到 62.2%　mIoU 的評分。FCN 雖然具有較高的 mIoU，但同時伴隨著龐大的計算量。

近年來，語義分割任務的成功有賴於大型標記資料集的開源，其中較知名的有 Camvid 資料集 [2]、Cityscapes 資料集 [3]、MSCOCO 資料集 [16]與 Pascal VOC 2012 資料集 [5]等。有關於語義分割的研究，基本上可分成以下幾個類型：（1）基於編碼器-解碼器之結構，其中比較著名的語義分割網路有 FCN [18]、SegNet [1]與 Fast-SCNN [20]等，其在 PaperWithCode Benchmarks（https://paperswithcode.com/sota/semantic-segmentation-on-ade20k）上有關 Cityscapes 資料集的 mIoU 分數分別為 65.3%、57.0%與 68.0%等。（2）基於注意力機制之結構，比較著名的方法有 PSANet [33]、CAA [13]與 MultiScale Spatial Attention [25]等，其在前述 Benchmarks 上有關 Cityscapes 資料庫 mIoU 分數分別為 81.4%、82.6%與 86.2%，其中文獻 [25]結合多尺度架構，目前取得第一的佳績。由此可見，植基於注意力機制與多尺度架構為未來語義分割研究的趨勢。

在單眼深度估測（Monocular depth estimation）的研究上，比較重要的資料庫包含有 KITTI [6, 7]、Make3D [26]與 NYU-Depth V2 [28]等。近年來，有關深度估計方法，如運動結構（Structure from Motion; SfM）以及立體視覺匹配（Stereo vision matching），都是建立在多視點的特徵對應上

（Feature correspondences）。有關於深度估測的方法，基本上可分成：（1）基於幾何的方法：透過幾何約束，從幾幅影像中恢復場景的 3D 結構，代表的方法有 SfM [31]，其可透過影像序列間的特徵對應及幾何約束來處理稀疏特徵的深度估測問題。因此，前述方法在深度估測的準確性，在很大程度上，與精確的特徵匹配和高品質的影像序列有關。（2）基於感測器的方法：有關於深度感測器，如 RGB-D 相機和光達（LiDAR），能夠直接擷取影像的深度資訊。RGB-D 相機能夠直接擷取 RGB 影像的像素級密集深度圖，其缺點為有限的測量範圍與光照敏感性。在無人自駕車應用上，光達 [32]是比較常用的方法，唯其僅能產生稀疏的 3D 地圖。（3）基於深度學習的方法：這是目前最流行的深度估測方法，在 KITTI Benchmarks （ http://www.cvlibs.net/datasets/kitti/eval_depth.php?benchmark=depth_prediction）的評分排行榜上，ViP-DeepLab [21]在 SILog 的評分指標上排行第 2。ViP-DeepLab 是一個深度模型，其提出主要用來試圖解決視覺中長期存在且具挑戰性的逆投影問題（Inverse projective problem），透過建模可從透視影像序列中恢復點雲，同時為每個點提供深度資訊。

二、研究方法

本文所提的深度神經網路如圖 1 所示，在所提的網路架構中總共包含有 6 個主要的卷積區塊，文中用 Stage 來表示。對於同一個 Stage，每個卷積層輸出的特徵圖具有相同的大小和通道數。在 1-6 的 Stage 中，它們包含 2-2-2-2-2-1 層的卷積區塊（Conv2D block），輸出通道的數量分別是 32-64-128-256-512-1024。

在本文中，所有卷積層都使用帶有可學習加權參數的卷積核。池化層使用 MaxPooling 來縮小輸出特徵圖的大小。在卷積層之後，應用批次正規化（Batch normalization; BN）來歸一化卷積層輸出的數據，以避免在反向傳播中出現梯度消失的現象，然後再使用 ReLU（Rectified Linear Unit）活化函數，其可以保持正值不變，但會將負值設為 0。

現在，我們將注意力轉向 Decoder 網路的細節，其中每個 Stage 對應於 Encoder 網路的相同 Stage。在 Decoder 網路中，每個卷積層表示為 DC-Conv-m-n，其中 DC，m 和 n 分別表示 Decoder、Stage 和 Layer。對於語義分割結構的設計，大多數編碼器網路都是相同的。唯一的區別在於解碼器網路架構。在本文中，我們修改 SegNet [1]的 Decoder 網路，同時引入跳躍連接的架構。這個想法的靈感主要來自 Lin 等人提出的特徵金字塔網路 [15]。該文確認了使用跳躍式連接結構時像素準確度（Pixel accuracy）具有較好的結果。

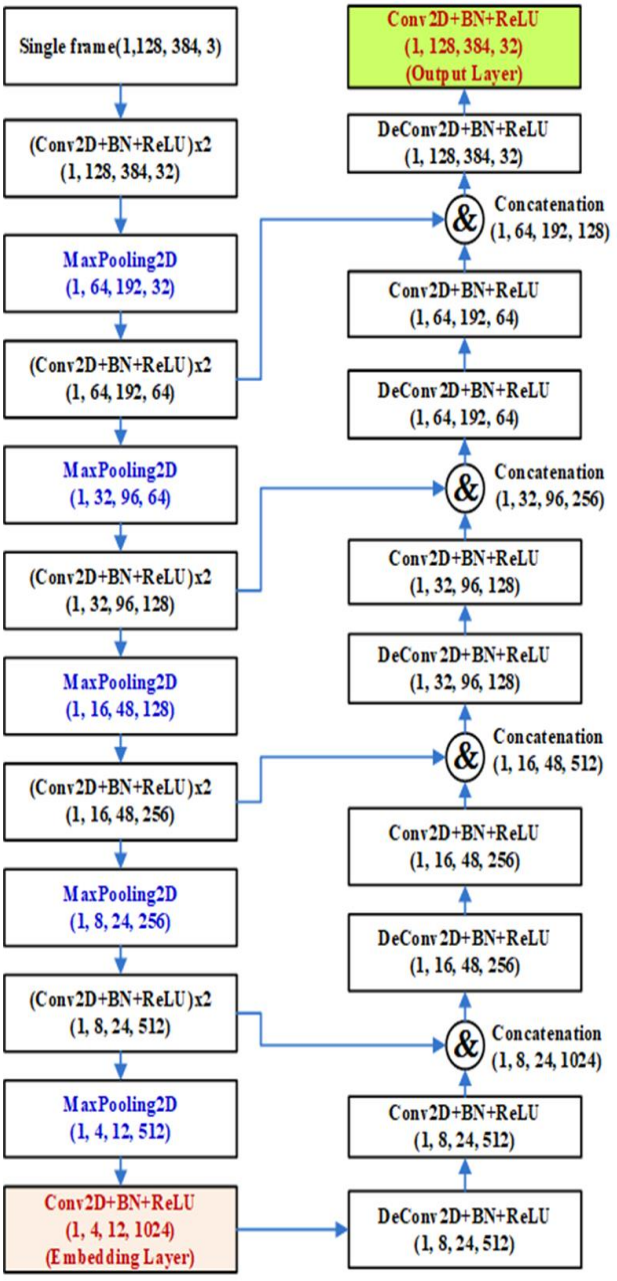

圖 1. 本文所提具有對稱 Encoder 和 Decoder 語義分割的網路架構。在圖中，×2 表示重複 2 次的卷積區塊。在小括號中(Batch, H, W, C)，其分別表示批次量大小，影像的高度、寬度以及通道的數量等

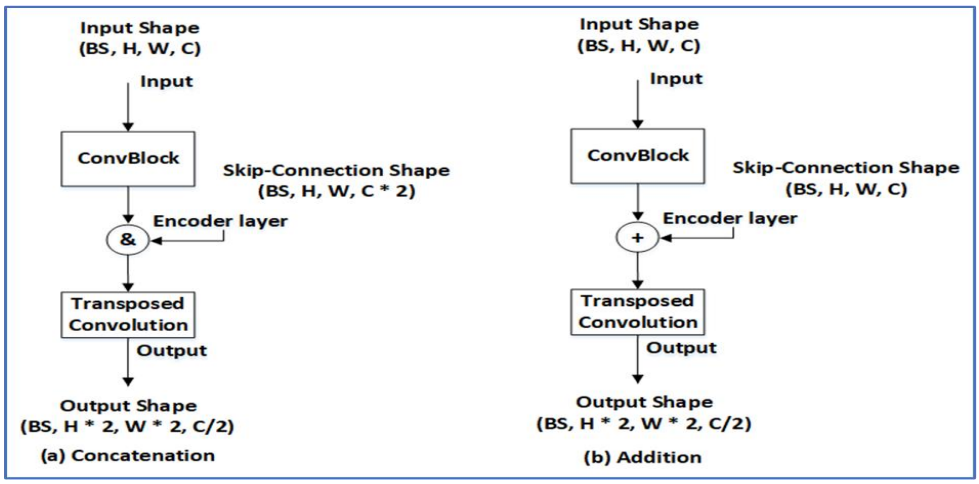

圖 2. 本文解碼器跳躍連接結構示意圖，其中(a)為串接方法、(b)為相加方法

為了更清楚描述文中所提跳躍連結的細部結構，我們以第 4 個 Stage 為例來進行說明。首先，我們在 Encoder 網路中選擇第 4 個 Stage 的最後一個卷積層，亦即 EC-Conv-4-3，因為在同一個 Stage 中最深的卷積層可以提取最具辨識度的特徵。然後，我們在 Decoder 網路中選擇相應的卷積層，亦即 DC-Conv-4-3。最後，再將這兩個層進行跳躍連接，如圖 2 所示。然後，再進行特徵圖放大以產生 Upsampling-3 層。

本文在語義分割解碼器的跳躍連接處加入注意力機制，其如圖 3 所示。在圖中，特徵圖影像 X（維度：BSxHxWxC）為主幹網路 Stage-3 Layer-3 或 Stage-4 Layer-3（EC-Conv-3-3 或 EC-Conv-4-3）之輸出特徵圖、Y（維度：BSxHxWxC）為經注意力機制區塊轉換後之輸出圖，其大小尺寸與 X 相同，其中 BS 為批次量大小，H 與 W 分別為特徵圖的高與寬，C 為特徵圖之通道數量。首先說明注意力機制之設計理念：變異數（Variance）與共變異數（Covariance）是統計學與機器學習中常用的統計量，其中變異數用來衡量隨機變數與平均值間的平方偏差量，然而共變異數則是用來衡量兩個隨機變數間之相似性。基於此，隨機變數間的分佈愈相似，共變異數就愈大；相反地，兩者間的相似性愈低，共變異數就愈小。在本文中，我們可將特徵圖中的每一個像素點視為一個隨機變數。因此，針對任一像素點（令為目標點）與所有其它像素點可計算其配對共變異數（Pairing covariances），例如 (x_1, x_2) 之配對共變異數為 $(x_1-\mu)(x_2-\mu)$。假設 $\mathbf{X}$ 為輸入的特徵圖，首先將 $\mathbf{X}\in\Re^{H\times W}$ 之形狀重新調整為 $\mathbf{a}\in\Re^{N\times 1}$，其中 $N=H\times W$，

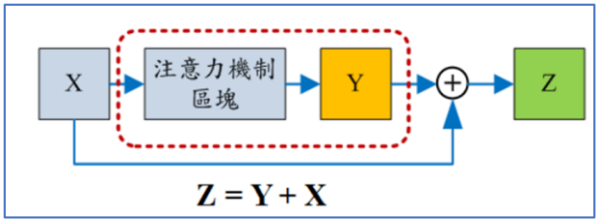

圖 3. 本文在編、解碼器的跳躍連接中加入注意力機制區塊

H 與 W 分別表示特徵圖 $\mathbf{X}$ 的高與寬。令 $\mathbf{a}=\mathbf{b}=\mathbf{c}=\begin{bmatrix}x_1 & x_2 & \cdots & x_N\end{bmatrix}^T$，並令μ為其平均值，因此共變異數 $\mathbf{Cov}_{N\times N}=(\mathbf{a}-\mu)(\mathbf{b}-\mu)^T$，進一步可計算注意力機制特徵圖為 $\mathbf{d}_{N\times 1}=\mathbf{Cov}_{N\times N}\cdot\mathbf{c}_{N\times 1}\Rightarrow\mathbf{Y}_{H\times W}$，最終特徵圖為原特徵圖與注意力機制特徵圖相加 $\mathbf{Z}=\mathbf{Y}+\mathbf{X}$。

三、結果與討論

本文實驗系統採用 Linux Ubuntu 18.04，開發環境為 Python 3.7.0，安裝的函式庫 TensorFlow 2.3.0 和 Opencv-python 3.2.0.8。本系統在 NVIDIA DGX Station 上進行實驗，該工作站配備有 Intel Xeon E5-2698 v4 2.2 GHz 處理器和 256GB 記憶體。DGX Station 配備有 4 張 Tesla V100 的 GPU 卡，每張 GPU 卡有 32GB 的記憶體。因此，全部 GPU 的記憶體為 128GB。

本文在 Cityscapes 資料集 [3]上進行所提深度神經網路在語義分割上之性能評估。而深度估測方面則於 KITTI 資料集 [6,7]上進行訓練及評估。本資料庫有大量的道路行

車紀錄檔，並包含對應感測器所得之深度真實值。在語義分割方面本文採用 mIoU（Mean intersection over union）的評估指標（Metrics），其常用於評估影像中物體分割的效能。至於 IoU 這種指標，則是用來評估影像中各個類別的分割效能。而深度估測評估度量則是採用 RMSE（Root mean square error）及準確性（Accuracies）

在本文中，我們使用 TensorFlow 來實現本文所提的深度神經網路架構。本文所提架構在 Cityscapes 資料集上進行訓練與測試。語義分割評估常用度量為 Pixel accuracy (PA) 及 mIoU 評分分別定義如公式（1）及（2）。

$$\mathrm{PA} = \frac{\sum_{i=0}^{C} p_{ii}}{\sum_{i=0}^{C}\sum_{j=0}^{C} p_{ij}}, \tag{1}$$

$$mIoU = \frac{1}{C+1}\sum_{i=0}^{C}\frac{p_{ii}}{\sum_{j=0}^{C} p_{ij} + \sum_{j=0}^{C} p_{ji} - p_{ii}}, \tag{2}$$

其中 C 是要預測的總類別數。由於背景也需要考慮進來，因此總類別數將增加為 C+1，這由公式（1）和（2）可以看出來，因為他們的下標索引計數為：i = 0 to C 以及 j = 0 to C。公式（1）與（2）中，符號p_{ii}表示該像素屬於第 i 個類別，且被識別為第 i 類，因此它是真陽性（True positive）；符號p_{ij}表示像素屬於第 i 個類別，但卻被錯誤地辨識為第 j 個類別，故其屬於偽陰性（False negative）；符號p_{ji}則是將第 j 個類別錯誤地標示為第 i 個類別，故其屬於偽陽性(False positive)。

本文為了評估深度估測網路的性能，本文採用文獻 [4] 中所提的評估方法，該評估方法有以下五個評估指標：RMSE、RMSE log、Abs Rel、Sq Rel 及 Accuracies，其分別定義如下：

$$RMSE = \sqrt{\frac{1}{N}\sum_{i\in I}\left\|d_i - d_i^*\right\|^2} \tag{3}$$

$$RMSE\ \log = \sqrt{\frac{1}{N}\sum_{i\in I}\left\|\log(d_i+1) - \log(d_i^*+1)\right\|^2} \tag{4}$$

$$Abs\ \mathrm{Rel} = \frac{1}{N}\sum_{i\in I}\frac{\left|d_i - d_i^*\right|}{d_i^*} \tag{5}$$

$$Sq\ \mathrm{Rel} = \frac{1}{N}\sum_{i\in I}\frac{\left\|d_i - d_i^*\right\|^2}{d_i^*} \tag{6}$$

$$Accuracy = \%\ of\ d_i\ s.t.\ \max\left(\frac{d_i}{d_i^*}, \frac{d_i^*}{d_i}\right) = \delta < thr \tag{7}$$

其中 d_i 與 d_i^* 分別表示影像深度的預測值與真值，I 為圖像，N 是圖像的總點數，thr 分別採用 1.25、1.25^2 及 1.25^3。以上指標主要用於評估影像深度真實值（Ground Truth）與預測值（Predicted Values）間接近的程度，其中 RMSE、RMSE log、Abs Rel 及 Sq Rel 等指標的值愈小代表深度網路的估測性能愈好；反之，Accuracy 指標是愈大愈好。

表 1 顯示在深度神經網路有否加入注意力機制對於語義分割效能之影響。由表中可知，當僅加入一層的注意力機制區塊較優於不加入注意力區塊之深度神經網路。這可從圖 4(b)與圖 4(c)中看出加入注意力機制區塊的語義分割效能是優於沒有加入注意力區塊的。同時，由表中亦可看出，當加入更多層的注意力機制區塊反而會劣化語義分割效能。

在語義分割方面，本文所提出的架構在 Cityscapes 資料集上進行了訓練與測試。本文提出的深度網路估測結果與相關文獻進行了比較，由表 2 可看出本文所提之深度神經網路架構在語義分割的各項評估結果優於文獻 [1, 18, 20]。

在深度估測方面，本文所提出的架構在 KITTI 資料集上進行了訓練與測試。我們提出的深度網路估測結果與相關文獻進行了比較，由表 3 可看出本文所提之深度神經網路架構在深度估測的各項評估結果都優於文獻 [4, 15]。

表 1. 針對深度神經網路架構中跳躍連接層是否加入注意力機制在 Cityscapes 資料集的 mIoU 和 Pixel Accuracy 的評分結果

Method	mIoU	Pixel Accuracy
No attention	0.7958	0.8856
Attention/Stage 4	**0.7961**	**0.8884**
Attention/Stage 3 & 4	0.6857	0.8115

註：粗體字表示該欄位中最佳的數值

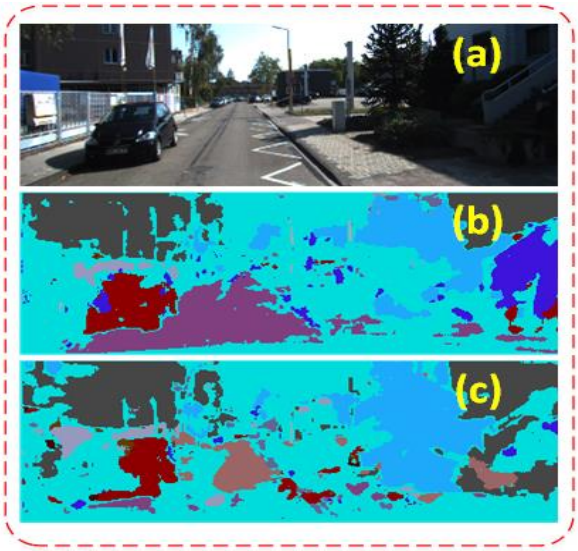

圖 4. 本文所提架構在解碼器增加注意力機制在語義分割方面的視覺結果比較：（a）原圖、（b）加入注意力機制之語義分割圖及（c）無注意力機制之語義分割圖，顯示加入注意力機制可以改善語義分割效能

最後本文在車輛與前方障礙物距離估測方面，從語義分割影像中取得分割之目標物，再與深度影像中取得相應位置之深度數值由小到大進行排序，取得前 20 百分位的深度數值作為該物體之距離估測數值，其如圖 5 所示，從圖 5(a)中可以看到本文所提方法能有效地估測出本車與前方障礙物間之距離。

表 2. 本文所提方法與現代語義分割方法在 mIoU 評分比較

Approach	mIoU（%）
Proposed (Attention/Stage 4)	**79.6**
Fast-SCNN [20]	68.0
FCN [18]	65.3
SegNet [1]	57.0

註：粗體字表示該欄位中最佳的數值

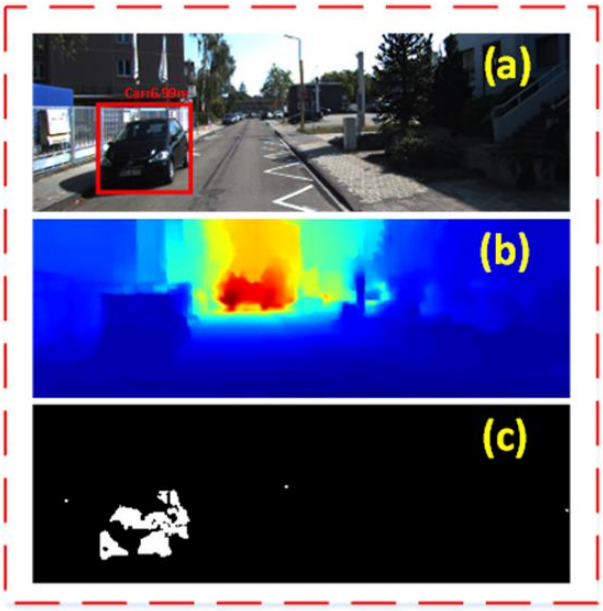

圖 5. 車輛與前方目標物(或障礙物)間之距離估測圖，（a）為原圖、（b）為深度影像圖、（c）為目標分割二值影像圖

表 3. 本文所提方法與相關文獻在深度估測效能之比較，測試的資料集為 KITTI Dataset

		Lower is better		Higher is better				
Approach	Depth	RMSE	RMSE(log)	ARD	SRD	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
[4] Coarse	0 – 80m	7.216	0.273	0.194	**1.531**	0.679	0.897	0.967
[4] Coarse+Fine	0 – 80m	7.156	0.270	0.190	1.515	0.692	0.899	0.967
[15] DCNF-FCSP	0 – 80m	7.046	--	**0.217**	--	0.656	0.881	0.958
Proposed	0 – 80m	**4.879**	**0.231**	0.158	1.101	**0.784**	**0.933**	**0.973**

註：粗體字表示該欄位中最佳的數值

四、結論

本文提出了一種對稱式 Encoder 和 Decoder 的深度神經網路架構。本文所提出的網路架構，在深度估測方面採用 KITTI 資料集，在語義分割則是採用 Cityscapes 資料集來分別進行測試。實驗結果顯示，本文所提障礙物距離估測方法的可行性。本文所提出的網路架構與其他相似的深度估測網路架構，在相同的訓練及測試條件下，本文提出的模型在準確率方面也有不錯的表現。在未來的工作中，將研究不同的解碼器架構以及更強健的障礙物偵測方法，以達成目標物的距離估測，同時持續改善本文所提深度估測網路之準確度。

誌謝

本文承科技部計畫「應用深度學習於行車記錄器中深度影像於障礙物偵測與距離估算，MOST 108-2221-E-212-012」補助完成，在此表達十分感謝之意。

參考文獻

1. Badrinarayanan, V., A. Kendall and R. Cipolla (2017) SegNet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12), 2481-2495.
2. Brostow, G. J., J. Fauqueur and R. Cipolla (2009) Semantic object classes in video: A high-definition ground truth database. *Pattern Recognition Letters*, 30(2), 88-97.
3. Cordts, M., M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth and B. Schiele (2016) The cityscapes dataset for semantic urban scene understanding. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV.
4. Eigen, D., C. Puhrsch and R. Fergus (2014) Depth map prediction from a single image using a multi-scale deep network. arXiv preprint arXiv:1406.2283.
5. Everingham, M. and J. Winn (2012) The Pascal Visual Object Classes Challenge 2012 (VOC2012) Development Kit, Retrieved April 01, 2021，from https://pjreddie.com/media/files/VOC2012_doc.pdf.
6. Geiger, A., P. Lenz and R. Urtasun (2012) Are we ready for autonomous driving? the kitti vision benchmark suite. Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI.
7. Geiger, A., P. Lenz, C. Stiller and R. Urtasun (2013) Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11), 1231-1237.
8. Hariharan, B., P. Arbelaez, R. Girshick and J. Malik (2015) Hypercolumns for object segmentation and fine-grained localization. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA.
9. He, K., G. Gkioxari, P. Dollár and R. Girshick (2017) Mask R-CNN. Proceedings of the IEEE International Conference on Computer Vision, Venice.
10. Hochreiter, S. and J. Schmidhuber (1997) Long short-term memory. *Neural computation*, 9(8), 1735-1780.
11. Horn, B., B. Klaus and P. Horn (1986) *Robot vision*. MIT .
12. Krizhevsky, A., I. Sutskever and G. E. Hinton (2012) ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 1097-1105.
13. Huang, Y., D. Kang, W. Jia, X. He and L. Liu (2021) Channelized axial attention for semantic segmentation. arXiv preprint arXiv:2101.07434.
14. Kuhnert, K. D. and M. Stommel (2006) Fusion of stereo-camera and pmd-camera data for real-time suited precise 3d environment reconstruction. IEEE/RSJ International Conference on Intelligent Robots and Systems, Beijing.
15. Lin, T. Y., P. Dollár, R. Girshick, K. He, B. Hariharan and S. Belongie (2017) Feature pyramid networks for object detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, Hawaii.
16. Lin, T. Y., M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár and C. L. Zitnick (2014) Microsoft COCO: Common objects in context. Proceedings in European Conference on Computer Vision, Zurich.
17. Liu, F., C. Shen, G. Lin and I. Reid (2015) Learning depth from single monocular images using deep convolutional neural fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(10), 2024-2039.
18. Long, J., E. Shelhamer and T. Darrell (2015) Fully convolutional networks for semantic segmentation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA.
19. Milgram, P., H. Takemura, A. Utsumi and F. Kishino (1995) Augmented reality: A class of displays on the reality-virtuality continuum. *Telemanipulator and Telepresence Technologies, International Society for Optics and Photonics*, 2351, 282-292.
20. Poudel, R. P. K., S. Liwicki and R. Cipolla (2019) Fast-SCNN: fast semantic segmentation network. arXiv preprint arXiv:1902.04502.
21. Qiao, S., Y. Zhu, H. Adam, A. Yuille and L. C. Chen (2021) ViP-DeepLab: Learning visual perception with depth-aware video panoptic segmentation, arXiv preprint arXiv:2012.05258.

22. Redmon, J., S. Divvala, R. Girshick and A. Farhadi (2016) You only look once: Unified, real-time object detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, Nevada.
23. Ren, S., K. He, R. Girshick and J. Sun (2015) Faster R-CNN: Towards real-time object detection with regon proposal networks. *Advances in Neural Information Processing Systems*, 91-99.
24. Ribler, R. L., J. S. Vetter, H. Simitci and D. A. Reed (1998) Autopilot: Adaptive control of distributed applications. Proceedings of the Seventh International Symposium on High Performance Distributed Computing (Cat. No. 98TB100244), Chicago, Illinois.
25. Sagar, A. and R. Soundrapandiyan (2020) Semantic segmentation with multiscale spatial attention for self driving cars. arXiv preprint arXiv:2007.12685.
26. Saxena, A., M. Sun and A. Y. Ng (2008) Make3D: Learning 3D scene structure from a single still image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(5), 824-840.
27. Sermanet, P., D. Eigen, X. Zhang, M. Mathieu, R. Fergus and Y. LeCun (2013) Overfeat: Integrated recognition, localization and detection using convolutional networks. arXiv preprint arXiv:1312.6229.
28. Silberman, N., D. Hoiem, P. Kohli and R. Fergus (2012) Indoor segmentation and support inference from rgbd images. Proceedings of the 12th European Conference on Computer Vision, Florence.
29. Simonyan, K. and A. Zisserman (2014) Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.
30. Szegedy, C., W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke and A. Rabinovich (2015) Going deeper with convolutions. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA.
31. Vijayanarasimhan, S., S. Ricco and C. Schmid (2017) SfM-Net: Learning of Structure and Motion from Video, arXiv preprint arXiv:1704.07804v1.
32. Yoneda, K., H. Tehrani, T. Ogawa, N. Hukuyama and S. Mita (2014) Lidar scan feature for localization with highly precise 3D map. IEEE Intelligent Vehicles Symposium Proceedings, Dearborn, MI.
33. Zhao, H., Y. Zhang, S. Liu, J. Shi, C. C. Loy, D. Lin and J. Jia (2018) PSANet: Point-wise spatial attention network for scene parsing. Proceedings of the European Conference on Computer Vision, Munich.

收件：110.04.04　修正：110.05.18　接受：110.06.22