Análisis de algoritmos de detección de objetos para la creación de un prototipo basado en la fusión de dos modelos de reconocimiento.

Analysis of object detection algorithms for the creation of a prototype based on the merger of two recognition models.

Miguel Ángel Quiroz Martínez[1,*]; Cristhian Paul Pangay Zambrano[1]; Kevin Johan Pérez Macías[1]
[1]Universidad Politécnica Salesiana, Ecuador.
{mquiroz@ups.edu.ec, cpangay@est.ups.edu.ec, kperezm2@est.ups.edu.ec}
Fecha de recepción: 27 de enero de 2019 ─ **Fecha de revisión:** 14 de marzo de 2019

Resumen: En el presente estudio se muestra una metodología experimental deductiva basada en redes neuronales para el reconocimiento de objetos con el uso de CNN. Nuestro objetivo es generar un prototipo el cual está basado en un mapa de características en combinación con RPN y propuesta de recorte en tronco que usa TNET para la detección 3D, dado por modelos de reconocimiento de objetos de la plataforma KITTI ,enfocados especialmente en AVOD y FPOINTNET, obteniendo una mayor precisión en objetos más pequeños que fácilmente son descartables por la nube de puntos proporcionados por el sensor laser 3d LIDAR HDL-64 pero no por el mapeado de características.
Palabras clave ─ RPN. FRUSTUN, KITTI, CNN, LIDAR.

Abstract: In the present study an experimental deductive methodology based on neural networks for the recognition of objects with the use of CNN is shown. Our goal is to generate a prototype which is based on a map of features in combination with RPN and a proposal of trimming on frustum that uses TNET for 3D detection, given by object recognition models of the KITTI platform, focused especially on AVOD and FPOINTNET, obtaining greater precision in smaller objects that can easily be discard by the cloud of points provided by the 3D LIDAR HDL-64 laser sensor but not by the feature map.
Keywords ─ RPN. FRUSTUN, KITTI, CNN, LIDAR.

Introducción

En el campo de la inteligencia artificial actualmente existe una competición en el desarrollo de algoritmos de reconocimiento de objetos, para su uso en la conducción autómata, estos algoritmos están basados en CNN (Kim, Y. ,2014), RNN (Zachary C. Lipton, John Berkowitz, 2015), DNN (Li, X., Hong, C 2013), entre otros. Las actuales tendencias en el desarrollo de algoritmos para el reconocimiento de objetos 3D están orientados a CNN (Kim, Y. ,2014), que permite clasificar el contenido de imágenes por fuerza bruta, que dominó el campo de clasificación de objetos desde el 2012 con ALEXNET (Iandola, F. N., Moskewicz, otros, 2016) a partir de entonces se han realizado múltiples investigaciones de redes convolucionales.

Pero el uso de CNN (Kim, Y. ,2014), carece de precisión en la detección de objetos más pequeños, en un marco de reconocimiento de objetos representa la dificultad de reconocimiento a una distancia mayor de 70 metros aproximadamente, los modelos de reconocimiento que usan diversos tipos de disparadores para el reconocimiento de objetos 2D como SSD o YOLO que se usan en modelos de reconocimiento como VOXELNET(Yin Zhou, Oncel Tuzel, 2017) y Fpointnet (Charles R, y otros, 2018), una aproximación diferente es la que usa MV3D (X. Chen, H. Ma, J. Wan, B. Li, and T. Xia, 2017) a nivel de capas o AVOD imitando el funcionamiento del reconocimiento por FPN.

Por ello nuestro objetivo es generar un prototipo a través de un diagrama que combine las funcionalidades utilizando un método deductivo y exploratorio de los modelos de la plataforma KITTI (A. Geiger, P. Lenz, and R. Urtasun) , mejorando la precisión de reconocimiento a través de un mapa de características y una red convolucional que detecte los objetos del mapa como RPN así usar esto para la creación de

[*] Maestro en Ingeniería con Especialidad en Sistemas de Calidad y Productividad

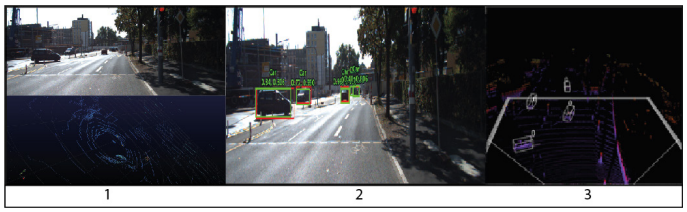
recortes en tronco para reconocer los objetos 3D que se encuentran en la nube de puntos proporcionada por el sensor láser HDL-64E LIDAR (B. Li, 2016).

Materiales y Métodos

Materiales

Esta investigación uso artículos relacionados con modelos de reconocimientos de objetos basados en CNN (Kim, Y, 2014) o variantes de ésta para analizar la arquitectura y funcionamiento de los algoritmos provenientes de la plataforma KITTI (A. Geiger, P. Lenz, and R. Urtasun) como AVOD (Ku, J. & Otros, 2017) F-PointNet (Charles R, y otros, 2018), VOXELNET (Yin Zhou, Oncel Tuzel, 2017), MV3D (X. Chen, H. Ma, J. Wan, B. Li, and T. Xia, 2017), entre otros, que utilizan la información proporcionada por el sensor láser 3d HDL-64E LIDAR (B. Li, 2016) la cual da nube de puntos que tiene un alcance máximo de 120 metros se usa en conjunto con las imágenes RGB para la fijación de objetos, en nuestro caso vehículos.

Figura 1. Estructura básica de un modelo de reconocimiento



Los algoritmos mencionados anteriormente están desarrollados en python con herramientas como tensorflow, mayavi, C++, etc, que en conjunto se usan para investigaciones de aprendizaje profundo y desarrollo de aplicaciones apoyadas con la arquitectura de cálculo paralelo de NVIDIA.

Estructura básica

Cada algoritmo se apoya en usos de diferentes en una metodología diferente, con distintos puntos de enfoque y resultados que varían, por ello determinamos una estructura con los puntos bases que se repite en estos algoritmos ,como lo demuestra la Figura 1, estos podrían ser interpretado como 3 pasos en donde primero a partir de una entrada de datos RGB y una nube de puntos aplica el uso de redes convolucionales para usar un algoritmo de detección de objetos que luego se usa al calcular la estimación del espacio representado por un cuadro delimitador del objeto 3D.

1. Método de Detección

Para realizar un reconocimiento de objetos con una mayor velocidad se usan detectores de tipo un solo disparo, al analizar diferentes tipos de detectores, se muestra en la Tabla 1, como el uso de SSD (C.-Y. Fu, W. Liu, A. Ranga, A. Tyagi, A. C. Berg, 2017) basado en detección de objeto por CNN (Kim, Y, 2014) que se encuentra en modelos algorítmicos como F-PointNet (Charles R, y otros, 2018), al analizar los resultados de su implementación se comprobó el descarte de objetivos a reconocer al momento de realizar las predicciones 3D.

El SSD en conjunto con otros sistemas de detección de un solo disparo como (YOLO YOLO (Redmon, J., Divvala, S, Girshick, R., Farhadi, 2016) tiene este tipo de problema o carencia al momento de reconocer objetos, y el uso de FPN (Tsung-Yi Lin y otros, 2017), muestra un mayor nivel de reconocimiento de objetos más lejanos aproximadamente a una distancia mayor a 70 metros, ya que funciona en conjunto con detectores de objetos.

Tabla 1. Detectores

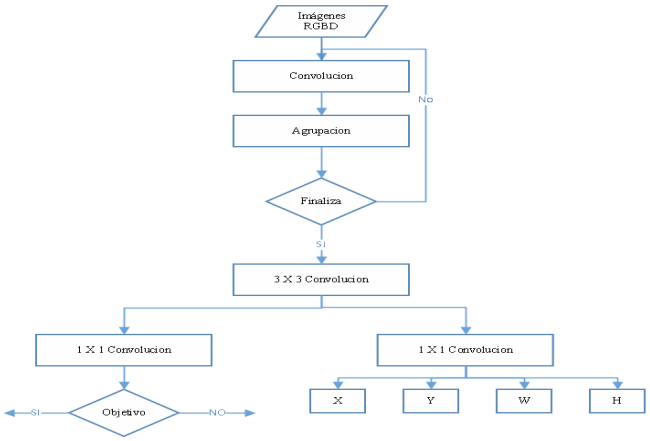
Tipo	*Descripción*
SSD	Detector de objetos basado en CNN que usa mapa de características y aplica 3X3 convoluciones para hacer predicciones
YOLO	Basado en CNN, realiza regresión lineal usando capas conectadas para hacer predicciones 7 X 7 X 2
FPN	Detector de características que funcionan con detectores de objetos

El uso de un detector de características en conjunto con uno de objetos para aumentar la precisión resulto en el uso de RPN (Song, S., Lichtenberg, S. P. & Xiao, J. ,2015) como detector a partir de un mapa de características como el que proporciona VGG16 (Simonyan, K. & Zisserman, A, 2014).

2. VGG16 junto a RPN para la identificación de objetos

La estructura proporcionada por VGG16 (Simonyan, K. & Zisserman, A, 2014) que funciona con 16 capas, y gracias a su efectividad llegó a un error de 7.5% en “ImageNet Large Scale Visual Recognition Competition 2012” (Simonyan, K. & Zisserman, A, 2014) permite crear un extractor de características reemplazando FPN (Tsung-Yi Lin y otros , 2017) permitiendo que trabaje junto a RPN (Song, S., Lichtenberg, S. P. & Xiao, J. ,2015) para detectar objetos más pequeños.

Figura 2. Estructura del funcionamiento de RPN [6]



Esta detección se realiza usando una estructura basada en CNN (Kim, Y, 2014) que permite generar propuestas para la región tomando el mapa de características y el resultado de la detección de objeto lo da la última capa convolucional como se explica en la figura 2.

3. PointNet para detección 3D

El modelo de detección de objetos Fpointnet (Charles R, y otros, 2018) en su arquitectura cuenta con propuesta de recorte en tronco, este proceso nos permite obtener un determinado campo de visión mediante datos suministrados como nube de puntos u objetos basado en el reconocimiento de obstáculos 2D. Una vez obtenidos los datos se procede a realizar un recorte 3D que se calcula usando *"n x c"*, en donde n es el número de objetivos y c el número de convoluciones, para centralizar el tronco.

La red de clasificación toma *n* puntos como entrada, luego realiza la transformación de la información obtenida, posterior a eso agrega características de punto por agrupación máxima, el resultado es la puntuación de clasificación. La red de segmentación es una extensión de la red de clasificación, concatena características y resultados globales y locales por puntaje usando puntos, posteriormente se combinará con el RPN (Song, S., Lichtenberg, S. P. & Xiao, J. ,2015). permite usar la vista 2D para reconocer todos los elementos que se encuentra dentro de un espectro de visión, y crear propuestas de recorte que aprovechara TNET (Li, X., Bing, L., Lam, W. & Shi, B. ,2018) y así procesar sus datos, realizar los cálculos, definiendo la caja delimitadora del objeto.

Metodología

Para esta investigación, se aplicó el método deductivo para determinar un algoritmo con mayor eficiencia y precisión. Para efectos del mismo se exploraron las estructuras de otros modelos, experimentando e implementando algoritmos como VOXELNET (Yin Zhou, Oncel Tuzel, 2017), AVOD(Ku, J. & Otros,2017) , FPOINTNET (Charles R, y otros, 2018) ,MV3D (X. Chen, H. Ma, J. Wan, B. Li, and T. Xia, 2017), usando la experiencia adquirida se identificó los puntos más relevantes de las estructuras de los algoritmos, a partir de una combinación de estas partes se obtiene un algoritmo con mayor precisión, menor uso de espacio en memoria logrando así un algoritmo más eficiente, nuestro trabajo consta de 4 fases que son:

1. En la primera fase, se realizó un análisis de las estructuras de los modelos de reconocimientos de objetos.
2. En la segunda fase se analizó los tipos de algoritmos de reconocimiento 2D para profundizar en el esquema de reconocimiento por RPN (Song, S., Lichtenberg, S. P. & Xiao, J., 2015).
3. En la tercera fase, se profundizo en el uso de TNET (Li, X., Bing, L., Lam, W. & Shi, B. ,2018) para la precisión 3D usando un esquema de recortes en forma de tronco alrededor del objeto.
4. En la cuarta fase realizó un prototipo mediante un diagrama que representa nuestra propuesta para realizar un algoritmo de reconocimiento de objetos.

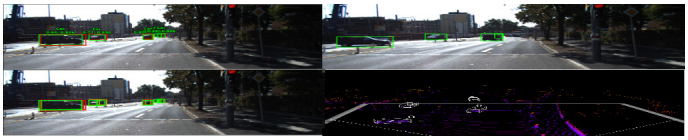
Resultados

Al realizarse un análisis entre los algoritmos como VOXELNET (Yin Zhou, Oncel Tuzel, 2017), AVOD(Ku, J. & Otros, 2017) , FPOINTNET (Charles R, y otros, 2018) ,MV3D (X. Chen, H. Ma, J. Wan, B. Li, and T. Xia, 2017) , además de la implementación de varios de estos modelos realizados con un procesador de 4.1 Ghz y una tarjeta gráfica de NVIDIA gtx1060 bajo un sistema operativo Ubuntu , el entrenamiento se realizó con los modelos proporcionados por la plataforma KITTI con el uso de la información proporcionada con el sensor láser HDL-64E LIDAR y el uso de imágenes RGBD.

Precisión entre modelos de reconocimiento

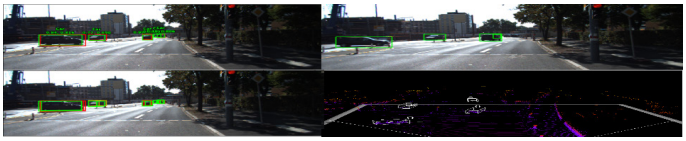
Se realizó una comparativa entre el uso de mapa de características junto a RPN (Song, S., Lichtenberg, S. P. & Xiao, J. ,2015) versus un reconocimiento en base CNN (Kim, Y. ,2014) de objetos, usando la misma entrada de datos RGBD y nube de puntos para iniciar el aprendizaje de los modelos.

Figura 3. Comparativa entre algoritmo de reconocimiento en ambiente 2D (RPN/CNN)



Mediante la implementación de estos modelos en función del tiempo de entrenamiento se pudo comprobar la efectividad de RPN (Song, S., Lichtenberg, S. P. & Xiao, J. ,2015) en comparativa de otros modelos basados en CNN (Kim, Y. ,2014), como se muestra en la Figura 3. En donde se puede apreciar el reconocimiento de objetivos más pequeños en este caso enfocado en vehículos este reconocimiento es realizado a partir de las imágenes RGBD.

Figura 4. Comparativa entre algoritmo de reconocimiento en ambiente 3D (RPN/ FPOINTNET)



De esta misma forma se analizó la detección de objeto 3D usando la nube de puntos proporcionada por el sensor láser LIDAR (Li, B. ,2016) comprobando que el modelo FPOINTNET (Charles R, y otros, 2018) logra una mayor precisión al momento de determinar la caja delimitadora del objeto como se muestra en la figura 4. También mostraba un uso menor en memoria al realizar el entrenamiento como se muestra en la Tabla 2.

Tabla 2. Peso y tiempo de entrenamiento Algoritmos con los resultados

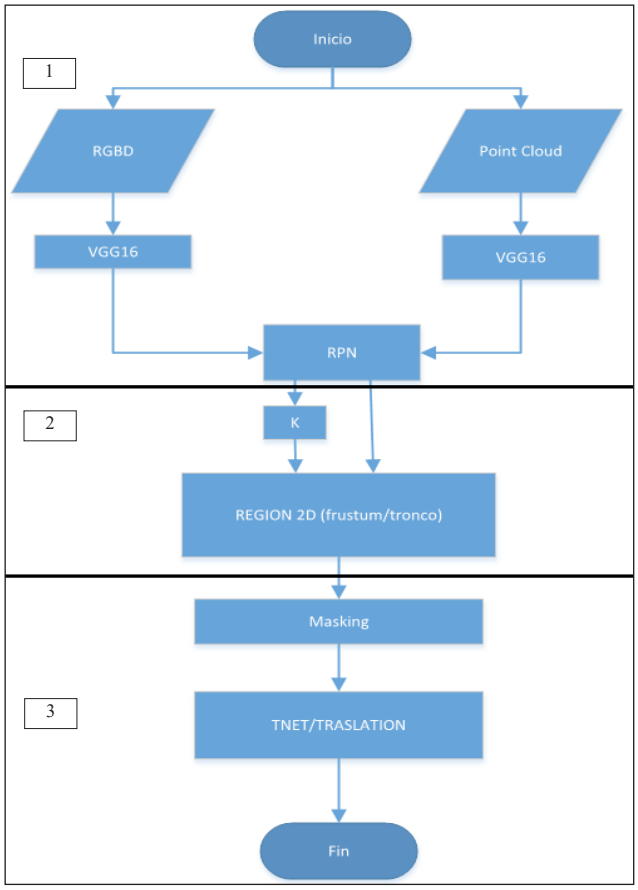
Tipo	*Peso* (aproximación)	*Tiempo de entrenamiento(horas)*
VOXELNET	100 Gb	48 (aproximación)
AVOD	75 Gb	38:24 (aproximación)
FPOINT	8 Gb	43:12 (aproximación)

Explicación del modelo

A partir del análisis realizado, se obtuvo como resultado un algoritmo utilizando técnicas de diagrama de flujo para el tratamiento que se puede ofrecer al campo de investigación de reconocimiento de objetos utilizando inteligencia artificial.

Nuestro modelo se basa en 3 fases: Primero la extracción de características, segundo: Recorte de objetos, tercero predicción de la caja delimitadora. Como se indica en la Figura 5.

Figura 5. Algoritmo propuesto



1. Extracción de características

Utiliza imágenes RGB y nube de puntos (proviene del sensor laser 3D y usa únicamente los puntos que pertenecen a la parte frontal del vehículo) para extraer las características utilizando VGG16 [15], como se muestra en la Figura 2, luego se fusionan las características 2D y 3D mediante un modelo pre-entrenado de RPN (Song, S., Lichtenberg, S. P. & Xiao, J. ,2015) que está inspirado en CNN (Kim, Y., 2014), esto permitirá reconocer los objetos que sean más pequeños.

RPN (Song, S., Lichtenberg, S. P. & Xiao, J. ,2015) aplica convoluciones de 3 × 3 esto es basado en CNN (Kim, Y. ,2014) que realiza un procedimiento por fuerza bruta en combinación con los extractores de características, seguidas de convolución separada de 1 × 1 para las predicciones de clase y la regresión de cuadro de límite logrando determinar el número

de predicciones que permite realizar los cálculos del centroide para realizar la centralización del objeto.

2. Recorte de objetos

Los objetos reconocidos por RPN (Song, S., Lichtenberg, S. P. & Xiao, J. ,2015) nos permite identificar dentro de la nube de puntos los troncos de los objetos que se identifican dentro del campo de visión, estos troncos estará definido por 8 puntos clave que determinaran el recorte de este que va desde donde se toma en cuenta la parte frontal del vehículo hasta la distancia límite del sensor laser HDL-64 LIDAR que es de 120 metros de largo con una reflectividad de 0.80 , esta información se utilizará para realizar el recorte de los objetos dentro de la nube de puntos.

3. Estimación del marco 3D

Para la creación del marco delimitador 3D se toman en cuenta características tales como: "centro (cx, cy, cz), tamaño (h, w, l) y ángulo de dirección () a lo largo del eje superior" Li, X., Bing, L., Lam, W. & Shi, B. ,2018. Una vez definidos los límites del marco se obtiene un enfoque residual para la estimación del centro del marco lo que proporciona la información para segmentar la instancia. PointNet (Qi, C. R., Su, H., Mo, K. & Guibas, L. J. ,2017) se encarga de clasificar el tamaño y el rumbo en las categorías predefinidas, y también los números residuales predecibles para cada categoría de forma simultánea se optimiza las tres redes involucradas, la segmentación de instancia 3D, TNET (Li, X., Bing, L., Lam, W. & Shi, B. , 2018 y estimación de caja amodal PointNet con pérdidas de tareas múltiples.

Mediante los puntos de objetos segmentados en las coordenadas de máscara 3D, se calcula el marco de limitación, que emplea el uso de TNET (Li, X., Bing, L., Lam, W. & Shi, B. ,2018) para la estimación del objeto que posteriormente se convierte en coordenadas, de tal modo que el centro predicho se convierta en el origen.

Conclusiones

Con el fin de crear un modelo de reconocimiento de objetos con una mayor precisión, se hizo un estudio de algoritmos que hagan el uso de redes neuronales, basado de su funcionamiento y arquitectura. De ser aplicada dentro de la industria automotriz y de esa forma apoyar este campo de desarrollo para la aplicación de estos modelos. A partir de la investigación realizada se tomó en consideración la estructura de reconocimiento que fusiona la visión de 2D/3D que utiliza AVOD y FPOINTNET, para la creación del prototipo propuesto, para lo cual se puede concluir lo siguiente: El reconocimiento de objetos tendrá una alternativa basado en una combinación de dos modelos de la plataforma de KITTI con el fin de mejorar el rendimiento y la precisión en la detección de objetos. El uso de la red neuronal se convierte en un aspecto fundamental en el reconocimiento de objetos la cual tiene como base el uso de CNN.

Bibliografía

- A. Geiger, P. Lenz, and R. Urtasun, (2012), Are we ready for autonomous driving? the kitti vision benchmark suite,IEEE Conference on Computer Vision and Pattern Recognition , 1,(1-6-8)
- Charles R. Qi,Wei Liu, Chenxia Wu, Hao Su, Leonidas J. Guibas , Stanford University, Nuro Inc. ,UC San Diego, 13 Apr 2018 , Frustum PointNets for 3D Object Detection from RGB-D Data,IEEE Conference on Computer Vision and Pattern Recognition ,1-2 , (2-3-4)
- C.-Y. Fu, W. Liu, A. Ranga, A. Tyagi, and A. C. Berg, (2017), Deconvolutional single shot detector,IEEE Conference on Computer Vision and Pattern Recognition,1,(1-2-3-4-5)
- Iandola, F. N., Moskewicz, M. W., Ashraf, K., Han, S., Dally, W. J. & Keutzer, K. (2016). SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <1MB model size.. CoRR, abs/1602.07360. (2)
- Kim, Y. (2014). Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882, (2,3,4,7,8)
- Ku, J., Mozifian, M., Lee, J., Harakeh, A. & Waslander, S. L. (2017). Joint 3D Proposal Generation and Object Detection from View Aggregation.. CoRR, abs/1712.02294. (2,3,8)
- Li, B. (2016). 3D Fully Convolutional Network for Vehicle Detection in Point Cloud.. CoRR, abs/1611.08069. (1,2)
- Li, X., Bing, L., Lam, W. & Shi, B. (2018). Transformation Networks for Target-Oriented Sentiment Classification.. In I. Gurevych & Y. Miyao (eds.), ACL (1) (p./pp. 946-956), Association for Computational Linguistics. ISBN: 978-1-948087-32-2
- Li, X., Hong, C., Yang, Y. & Wu, X. (2013). Deep neural networks for syllable based acoustic

modeling in Chinese speech recognition.. APSIPA (p./pp. 1-4), : IEEE. ISBN: 978-1-4799-2794-4

- Qi, C. R., Su, H., Mo, K. & Guibas, L. J. (2017). PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation.. CVPR (p./pp. 77-85), (2,5,6,8)
- Redmon, J., Divvala, S., Girshick, R., Farhadi, A, (2016) You only look once: Unified, real-time object detection, IEEE Conference on Computer Vision and Pattern Recognition,1, (1-5)
- Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L, (2015), Imagenet large scale visual recognition challenge, International Journal of Computer Vision, 1, (1-2-3-5-6)
- Simonyan, K. & Zisserman, A. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition. CoRR, abs/1409.1556. (4)
- Song, S., Lichtenberg, S. P. & Xiao, J. (2015). SUN RGB-D: A RGB-D scene understanding benchmark suite.. CVPR (p./pp. 567-576), IEEE Computer Society. ISBN: 978-1-4673-6964-0 (1,2,4,5,7)
- Tsung-Yi Lin, Piotr Dollar, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie, (2017), Feature Pyramid Networks for Object Detection, Computer Vision and Pattern Recognition, 1-2, (1-2-3-4-5-6-7)
- X. Chen, H. Ma, J. Wan, B. Li, and T. Xia, (2017), Multi-view 3d object detection network for autonomous driving. IEEE Conference on Computer Vision and Pattern Recognition, 1-3, (2-5- 6-11-12-13)
- Yin Zhou, Oncel Tuzel, VoxelNet, (2017), End-to-End Learning for Point Cloud Based 3D Object Detection,IEEE Conference on Computer Vision and Pattern Recognition ,1-2,(1-2-3)
- Zachary C. Lipton, John Berkowitz, (2015), A Critical Review of Recurrent Neural Networks for Sequence Learning, Machine Learning,1-2,(1-4)