ISSN 1870-4069

Reconstrucción de profundidad a partir de una sola imagen con perspectiva mediante redes neuronales completamente convolucionales

José E. Valdez Rodríguez, Hiram Calvo, Edgardo M. Felipe Riverón

Instituto Politécnico Nacional, Centro de Investigación en Computación, Ciudad de México, México

jvaldezr1000@alumno.ipn.mx, {hcalvo, edgardo}@cic.ipn.mx

Resumen. La reconstrucción de la profundidad a partir de una sola imagen ha sido una tarea difícil debido a la complejidad y la cantidad de indicios de profundidad que puede contener una sola imagen. Las Redes Neuronales Convolucionales (RNC) se han utilizado con éxito para reconstruir la profundidad de objetos en escenas generales; sin embargo, estos trabajos no se han adaptado para el problema particular de la reconstrucción de la profundidad a partir de la perspectiva. Nuestra propuesta se basa en construir una RNC computacional eficiente; nos centramos en Redes Neuronales Completamente Convolucionales (RNCC), cuyo entrenamiento requiere de una sola etapa. Comparamos nuestra RNC con el estado del arte actual en reconstrucción de profundidad, obteniendo mejoras en niveles globales para imágenes viales con perspectiva presente.

Palabras clave: Reconstrucción de profundidad, redes neuronales completamente convolucionales, RNCC, emparejamiento estereoscópico.

Reconstruction of Depth from a Single Perspective Image through Completely Convolutional Neural Networks

Abstract. The reconstruction of the depth from a single image has been a difficult task due to the complexity and the amount of depth indications that a single image can contain. Convolutional Neural Networks (RNC) have been used successfully to reconstruct the depth of objects in general scenes; however, these works have not been adapted for the particular problem of depth reconstruction from perspective. Our proposal is based on building an efficient computational RNC; we focus on Completely Convolutional Neural Networks (RNCC), whose training requires a single stage. We compared our RNC with the current state of the art in depth reconstruction, obtaining improvements in global levels for road images with present perspective.

Keywords: Depth Reconstruction, Completely Convolutional Neural Networks, RNCC, Stereoscopic Pairing.

1. Introducción

La reconstrucción de la profundidad a partir de una sola imagen (opuesto a las imágenes estereoscópicas) es una tarea difícil debido a la cantidad de señales de profundidad que una sola imagen puede contener, tales como sombras, perspectiva, imagen borrosa por el movimiento [12]. En este trabajo nos enfocamos en reconstruir la profundidad en imágenes cuya perspectiva es dominante; un ejemplo pueden ser imágenes de autopistas o calles. Este tipo de imágenes necesitan ser analizadas por sistemas autónomos tales como robots o automóviles los cuales solamente cuentan con una representación 2D para calcular su desplazamiento [4,14,10]. Diversos trabajos previos en reconstrucción de profundidad han abordado el problema general de estimar la componente de profundidad en una amplia gama de imágenes, centrándose en objetos contra un fondo sólido o complejo, o edificios de ciudades y personas.

Hasta donde sabemos, no hay trabajos específicos dedicados al problema de la reconstrucción de la profundidad utilizando RNCC. Varias arquitecturas de RNC en el estado del arte proponen redes de múltiples etapas que requieren entrenamiento separado para cada una de ellas, lo cual hace que el proceso de entrenamiento tome más tiempo. En este trabajo proponemos una arquitectura de Red Neuronal Completamente Convolucional (RNCC) capaz de reconstruir la profundidad a partir de una sola imagen, utilizando únicamente capas convolucionales, y por tanto requiere sólo una etapa de entrenamiento. La arquitectura propuesta en este trabajo es capaz de reconstruir la profundidad a nivel global y local. La organización de este trabajo se muestra a continuación: en la sección 2 se describe el trabajo relacionado en el cual está basada esta investigación, en la sección 3 se describe el método propuesto, en la sección 4 se muestran nuestros resultados, así como la comparación con el estado del arte, y finalmente en la sección 5 se muestran las conclusiones de este trabajo.

2. Trabajo relacionado

El problema de reconstruir profundidad a partir de una sola imagen ha sido atacado a través de diferentes métodos. Aunque el utilizar Redes Neuronales Convolucionales (RNC) se ha convertido en una de las mejores técnicas para resolver este problema, esta técnica sigue en etapa de desarrollo debido principalmente a la complejidad en el diseño de este tipo de redes neuronales. Uno de los primeros trabajos en utilizar esta técnica se puede ver en Eigen, Puhrsch y Fergus [6]. Ellos proponen el uso de dos RNC: La primera reconstruye la profundidad a nivel global y la segunda refina los detalles locales. Ellos establecen una RNC que

puede ser aplicada a la reconstrucción de profundidad y proponen una función de error.

Eigen y Fergus [5] utilizan tres RNCs entrenadas por separado. La primera reconstruye la profundidad a nivel global; la segunda trata de reconstruir la profundidad a nivel global a la mitad de resolución espacial de la imagen de entrada y la tercera refina los detalles a nivel local; además, proponen una nueva función de error para entrenar sus RNC.

Liu, Shen y Lin [18] presentan una RNC que toma como entrada una imagen pre-procesada basada en superpixeles. Ellos complementan con un método probabilístico, en el cual tratan de mejorar los resultados. Después, del mismo equipo, Liu, Shen, Lin y Reid [19] cambia la arquitectura de su RNC manteniendo la etapa de mejoramiento de la imagen.

Finalmente, Afifi y Hellwich [2] utilizan una sola RNC configurada con regresión para reconstruir la profundidad con su propia función de error.

En general, las arquitecturas discutidas anteriormente son similares, los cambios entre ellas están basados en la función de error, las funciones de activación entre las capas de las RNCs, el número de filtros por capa y el tamaño de los filtros. Dado que los autores de los trabajos relacionados utilizan diferentes conjuntos de datos para entrenar y probar sus RNC, se dificulta la comparación directa entre el desempeño nuestra propuesta y el trabajo relacionado no dejándonos otra opción que reimplementar alguno de las RNC para poder realizar una comparación adecuada. Dentro de los trabajos que se consideran, el trabajo que obtiene los mejores resultados en el estado del arte es el de Afifi y Hellwich [2]. Además, su arquitectura está basada en una RNC con un solo paso de entrenamiento.

A pesar de que su RNC es llamada completamente convolucional (fully convolutional), al utilizar una capa *upsample* y bloques residuales, consideramos que su RNC no es puramente convolucional. Los autores presentan resultados de reconstrucción de profundidad sobre imágenes de sillas con diferente fondo y perspectiva. Dado que nos interesa evaluar el desempeño de nuestra RNCC con un mismo conjunto de imágenes en las cuales la perspectiva es un factor importante, re-implementamos la RNC de Afifi y Hellwich [2] como se ve en la Figura 1.

En la Región A se puede observar la operación de reducción, la cual es realizada a través de cuatro bloques residuales y una capa convolucional. En la región B se observa la operación *upsample* la cual trata de recuperar el tamaño original de la imagen de entrada y ésta se realiza a través de dos bloques residuales y una capa convolucional. En la sección 3 se darán más detalles acerca de estas operaciones.

En la siguiente sección presentamos nuestra propuesta. Al igual que la RNC de Afifi y Hellwich, nuestra RNCC está basada en una sola RNC, pero la diferencia principal es que nuestra RNCC está compuesta solamente de capas convolucionales y no recupera el tamaño de la imagen original (*upsampling*) sobre la misma arquitectura de la RNCC. Nuestra arquitectura tampoco emplea bloques residuales.

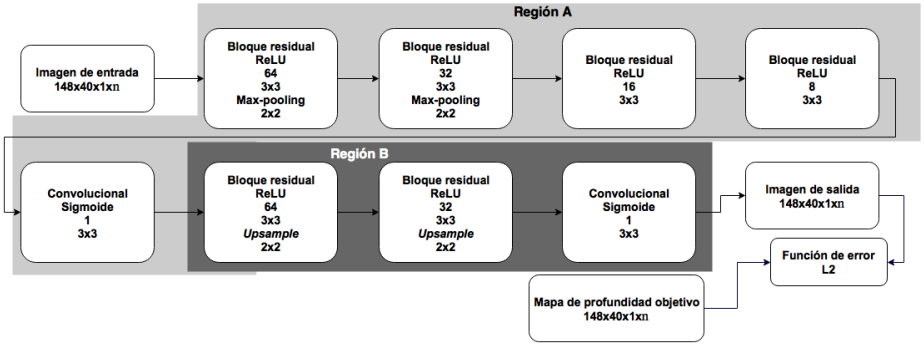

Fig. 1. Arquitectura de RNC de Afifi y Hellwich [2].

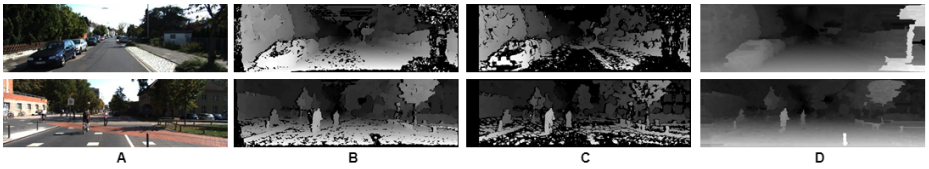

Fig. 2. Comparación entre los mapas de profundidad obtenidos con los diferentes algoritmos de emparejamiento estereoscópico, (A) Imágenes estereoscópicas, (B)*Semiglobal Matching*, (C)*Blockmatching method*, (D)*Efficient large-scale stereo matching method*.

3. Método propuesto

En esta sección presentamos la arquitectura de nuestra de Red Neuronal Completamente Convolucional (RNCC). Hasta donde sabemos, no hay conjuntos de datos públicos disponibles para entrenar nuestra RNCC por lo que se decidió crear un conjunto de datos a partir de la sección *Object tracking* de *The KITTI Vision Benchmark Suite* [7]. Este conjunto de datos contiene 15,000 escenas al aire libre brindadas como imágenes estereoscópicas. Este conjunto de datos no contiene los mapas de profundidad objetivo, por lo que hemos construido dicho mapa utilizando diferentes algoritmos de emparejamiento estereoscópico, tales como *Semiglobal stereo matching* [11] y *Blockmatching* [13].

En la Figura 2 podemos observar una breve comparación entre los algoritmos de emparejamiento estereoscópico probados para obtener el mapa de profundidad objetivo. Por último, hemos decidido utilizar el algoritmo *Efficient large-scale stereo matching* [8], debido a la calidad y la evaluación recibida por Menze y Geiger [20]. Este algoritmo recibe pares de imágenes estereoscópicas, como se describe en la Figura 3, y su resultado es el mapa de profundidad objetivo.

Nuestra RNCC fue construida a partir de capas Convolucionales [15] y *Max-pooling* [21]. Adicionalmente, agregamos un sesgo a cada capa de la RNCC. En la Figura 4 se puede observar la representación de cada capa en la RNCC.

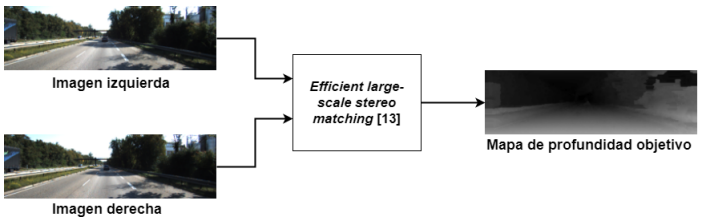

Fig. 3. Creación del mapa de profundidad objetivo.

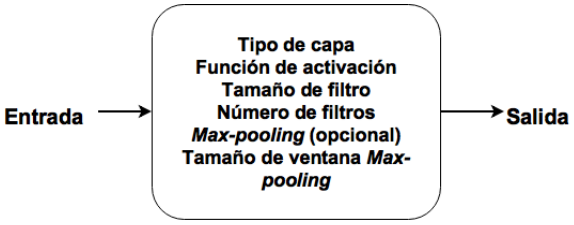

Fig. 4. Representación de una capa de la RNCC.

RNC propuesta. El diagrama a bloques de esta RNC se presenta en la Figura 5. Esta RNC está compuesta de cinco capas, de las cuales cuatro son capas convolucionales con función de activación *Rectified Linear Unit* (ReLU) [3] y la salida de la RNCC tenemos una capa convolucional con función de activación sigmoide para limitar la salida de la RNCC entre valores de 0 y 1. Esta red neuronal recibe como entrada la imagen izquierda de tamaño 147x37 pixeles en niveles de gris y como mapa de profundidad objetivo recibe el mapa de profundidad correspondiente a la entrada obtenido con el algoritmo de emparejamiento estereoscópico de tamaño 37x10 pixeles en niveles de gris. A la salida la RNCC nos entrega el mapa de profundidad estimado con un tamaño de 37x10 pixeles en niveles de gris. Este tamaño se debe a que se están utilizando capas max-pooling en las dos primeras capas convolucionales.

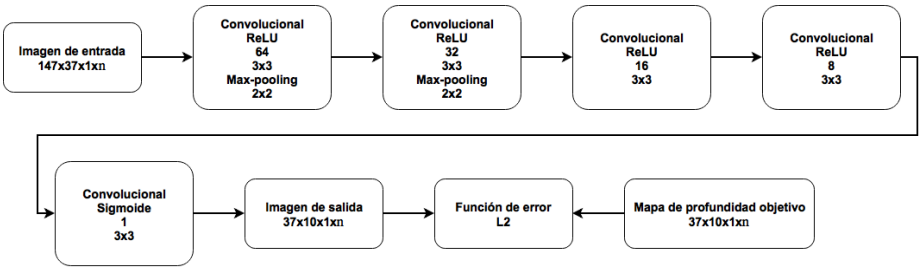

Fig. 5. Arquitectura de RNCC propuesta.

Función objetivo. Para entrenar nuestra RNCC necesitamos minimizar el error. Esta función mide la diferencia global entre el mapa de profundidad objetivo y la reconstrucción de profundidad dada por la RNCC. Utilizamos la norma *L2* como la función objetivo. La norma *L2* se puede calcular como se muestra en la ecuación 1.

$$L2 = \frac{1}{2n} \sum_{i=1}^{n} \|y(i) - y'(i)\|_2^2 , \tag{1}$$

donde:

y'(i) = Mapa de profundidad reconstruido,
y(i) = Mapa de profundidad objetivo,
n = Número de imágenes por lote.

Medidas de error. Al evaluar nuestro método utilizamos varias medidas de error para poder compararnos con el trabajo relacionado:

Raíz del Error Cuadrático Medio:

$$RECM = \sqrt{\frac{1}{|T|} \sum_{y' \epsilon |T|} (y - y')^2}. \tag{2}$$

Error Cuadrático Medio:

$$EMC = \frac{1}{|T|} \sum_{y' \epsilon |T|} (y - y')^2. \tag{3}$$

Raíz del Error Cuadrático Medio Logarítmico:

$$RECMLOG = \sqrt{\frac{1}{|T|} \sum_{y' \epsilon |T|} (log(y) - log(y'))^2}. \tag{4}$$

Raíz del Error Cuadrático Medio Logarítmico Invariante a la Escala:

$$RECMLOGSI = \frac{1}{|T|} \sum_{y' \epsilon |T|} (log(y) - log(y'))^2. \tag{5}$$

Diferencia Absoluta Relativa:

$$DAR = \frac{1}{|T|} \sum_{y' \epsilon |T|} \frac{|y - y'|}{y'}. \tag{6}$$

Diferencia Cuadrada Relativa:

$$DCR = \frac{1}{|T|} \sum_{y' \epsilon |T|} \frac{\|y - y'\|^2}{y'}. \tag{7}$$

En las fórmulas:

y' = Mapa de profundidad reconstruido,
y = Mapa de profundidad objetivo,
T = Número de pixeles en la imagen.

4. Experimentos y resultados

En esta sección se describirán los resultados de nuestra propuesta. Nuestra RNCC fue entrenada a través del algoritmo de Retropropagación [17] y Descenso de Gradiente Estocástico [16]. Utilizamos 1,000 épocas y un tamaño de lote de 40. Del conjunto de datos completo tomamos 12,482 imágenes, solamente del lado izquierdo con su respectivo mapa de profundidad objetivo para entrenar la RNCC, y 2,517 imágenes respectivamente para prueba.

El tamaño de las imágenes de entrada a la RNCC es de 147x37 pixeles y el tamaño de la imagen de salida es de 37x10 pixeles. Para recuperar el tamaño original de las imágenes y poder compararnos con resultados obtenidos en trabajos del estado del arte utilizamos Interpolación Bilineal [9]; con lo anterior tenemos $T = 370$ (sin recuperar el tamaño original) y $n = 40$ dado el tamaño del lote. Nuestra RNCC fue implementada en un entorno de trabajo de Python, *Tensorflow* [1] en el cual las RNCC pueden ser entrenadas en GPU para obtener un mejor desempeño.

Ambas arquitecturas de RNC se entrenaron en una GPU NVIDIA GTX 960; tomó dos días realizar el entrenamiento y menos de un segundo en reconstruir la profundidad cuando la RNC recibe una sola imagen. Como se explicó en la sección 2, se implementó la RNC propuesta por Afifi y Hellwich [2] dado que ellos solamente entrenan su RNC una vez; esta RNC también puede ser entrenada con la norma *L2* y aun así ellos obtienen mejores resultados que el resto del trabajo relacionado sin utilizar una RNC adicional. En la Figura 6 se puede observar una muestra de los resultados obtenidos con la implementación de [2].

En la Figura 7 Se presenta una muestra de los resultados obtenidos con nuestra RNCC.

Comparando cualitativamente nuestros resultados mostrados en la figura anterior, se puede observar que nuestra RNC es capaz de reconstruir la profundidad a nivel global. A nivel local, se puede observar que la RNC es capaz de mostrar más detalles que en propuestas anteriores. Por nivel global nos referimos a la imagen sin detalles pequeños y a nivel local nos referimos a los pequeños detalles que pueda contener la imagen, tales como automóviles, peatones, postes, etc. Para un análisis cuantitativo, en la Tabla 1 presentamos las medidas de error de nuestra propuesta y la RNC implementada mencionado anteriormente. Los errores REMC, EMC y DAR miden el error a nivel global de las imágenes, mientras que los errores REMCLOG, REMCLOGSI y DCR miden los errores a nivel local. Comparando nuestro método con el estado del arte, obtenemos mejores resultados en la mayoría de medidas del error a nivel global pero a nivel local se necesita mejorar.

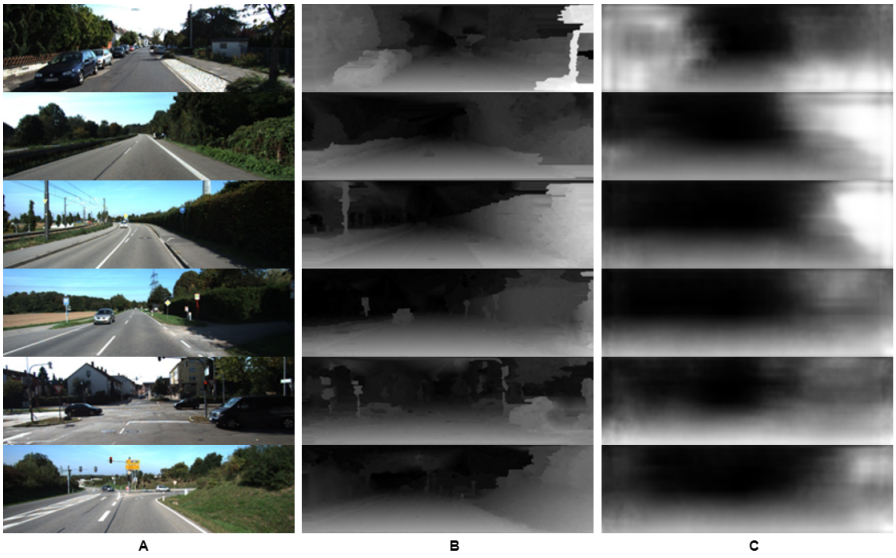

Fig. 6. Resultados de la RNC de Afifi and Hellwich [2] con nuestro conjunto de datos. (A) Imagen de entrada, (B) Mapa de profundidad objetivo y, (C) Salida de la RNC y'.

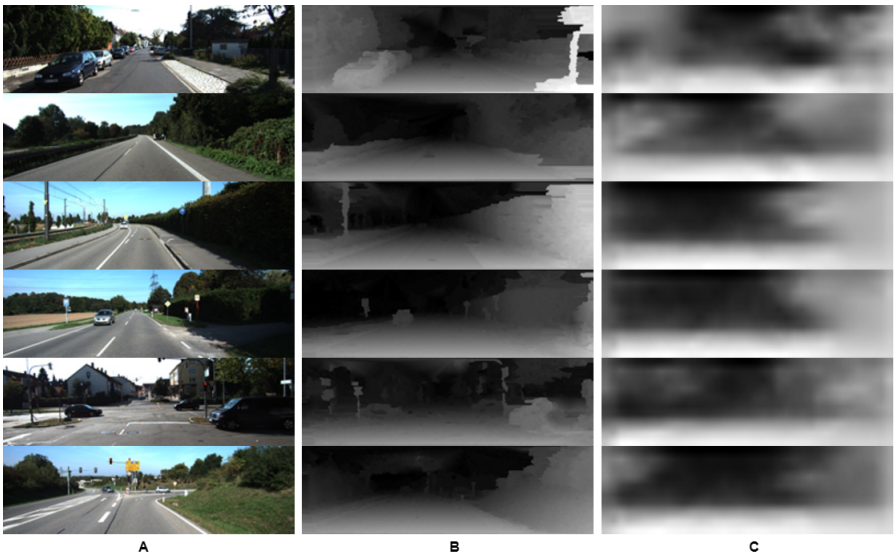

Fig. 7. Muestra de resultados obtenidos con nuestra RNCC. (A) Imagen de entrada, (B) Mapa de profundidad objetivo y, (C) Salida de la RNCC y'.

Tabla 1. Comparación de resultados entre nuestra propuesta y el estado del arte.

	RECM	ECM	RECMLOG	RECMLOGSI	DAR	DCR
Estado del arte [2]	0.1496	0.0260	0.3618	**0.2068**	**9.8658**	**45.1089**
RNC propuesta	**0.1301**	**0.0193**	**0.3317**	0.2269	11.1982	58.5867

5. Conclusiones y trabajo futuro

Hemos presentado una arquitectura de RNC capaz de reconstruir la profundidad utilizando solamente capas convolucionales, y por tanto, que requiere únicamente de entrenamiento en una sola etapa. Hemos adaptado un conjunto de datos existente de imágenes estereoscópicas para desarrollar un recurso orientado a probar la reconstrucción en profundidad de imágenes con perspectiva. Utilizamos este conjunto de datos para probar nuestra RNCC y comparamos su desempeño con el estado del arte actual en reconstrucción de profundidad. Encontramos que el uso de capas convolucionales ayuda a mejorar los resultados a nivel global. La perspectiva de la imagen toma un papel importante en la reconstrucción de la profundidad. Cuantitativamente, la profundidad local se puede estimar con nuestra RNCC, pero esto todavía necesita ser validado con otros conjuntos de datos.

Como trabajo futuro planeamos experimentar con diferentes tamaños de filtro, diferentes funciones de error y funciones de activación para mejorar aún más nuestra RNCC propuesta, así como el uso de una etapa de refinamiento de la imagen.

Agradecimientos. Agradecemos al Instituto Politécnico Nacional (SIP, COFAA, EDD, EDI and BEIFI), y CONACyT por su apoyo a esta investigación.

Referencias

1. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., Zheng, X.: TensorFlow: Large-scale machine learning on heterogeneous systems (2015), http://tensorflow.org/, software disponible en tensorflow.org
2. Afifi, A.J., Hellwich, O.: Object depth estimation from a single image using fully convolutional neural network. In: Digital Image Computing: Techniques and Applications (DICTA), 2016 International Conference on. pp. 1–7. IEEE (2016)
3. Arora, R., Basu, A., Mianjy, P., Mukherjee, A.: Understanding deep neural networks with rectified linear units. arXiv preprint arXiv:1611.01491 (2016)
4. Bills, C., Chen, J., Saxena, A.: Autonomous MAV flight in indoor environments using single image perspective cues. In: Robotics and automation (ICRA), 2011 IEEE international conference on. pp. 5776–5783. IEEE (2011)
5. Eigen, D., Fergus, R.: Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2650–2658 (2015)
6. Eigen, D., Puhrsch, C., Fergus, R.: Depth map prediction from a single image using a multi-scale deep network. In: Advances in neural information processing systems. pp. 2366–2374 (2014)

7. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2012)
8. Geiger, A., Roser, M., Urtasun, R.: Efficient large-scale stereo matching. In: Asian conference on computer vision. pp. 25–38. Springer (2010)
9. Gonzalez, W., Woods, R.E.: Eddins, digital image processing using MATLAB. Third New Jersey: Prentice Hall (2004)
10. Häne, C., Sattler, T., Pollefeys, M.: Obstacle detection for self-driving cars using only monocular cameras and wheel odometry. In: Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on. pp. 5101–5108. IEEE (2015)
11. Hirschmuller, H.: Stereo processing by semiglobal matching and mutual information. IEEE Transactions on pattern analysis and machine intelligence 30(2), 328–341 (2008)
12. Howard, I.P.: Perceiving in depth, volume 1: basic mechanisms. Oxford University Press (2012)
13. Konolige, K.: Small vision systems: Hardware and implementation. In: Robotics research, pp. 203–212. Springer (1998)
14. Kundu, A., Li, Y., Dellaert, F., Li, F., Rehg, J.M.: Joint semantic segmentation and 3D reconstruction from monocular video. Georgia Institute of Technology (2014)
15. LeCun, Y., Boser, B.E., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W.E., Jackel, L.D.: Handwritten digit recognition with a back-propagation network. In: Advances in neural information processing systems. pp. 396–404 (1990)
16. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE 86(11), 2278–2324 (1998)
17. LeCun, Y.A., Bottou, L., Orr, G.B., Müller, K.R.: Efficient backprop. In: Neural networks: Tricks of the trade, pp. 9–48. Springer (2012)
18. Liu, F., Shen, C., Lin, G.: Deep convolutional neural fields for depth estimation from a single image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5162–5170 (2015)
19. Liu, F., Shen, C., Lin, G., Reid, I.: Learning depth from single monocular images using deep convolutional neural fields. IEEE transactions on pattern analysis and machine intelligence 38(10), 2024–2039 (2016)
20. Menze, M., Geiger, A.: Object scene flow for autonomous vehicles. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2015)
21. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: European conference on computer vision. pp. 818–833. Springer (2014)