ISSN 1870-4069

Red neuronal convolucional de baja latencia para la estimación de profundidad monocular

Edgar Rodrigo López Silva, Luis Rogelio Román Rivera,
Jesús Carlos Pedraza Ortega, Marco Antonio Aceves Fernández,
Juan Manuel Ramos Arreguín

Universidad Autónoma de Querétaro,
Facultad de Ingeniería,
México

edgar.lopez06@outlook.com,
luis.rogelio.roman.rivera@gmail.com, caryoko@yahoo.com,
marco.aceves@gmail.com, jsistdig@yahoo.com.mx

Resumen. La estimación de profundidad monocular se ha convertido recientemente en una de las técnicas emergentes y de creciente interés para la obtención de la estructura espacial de una escena. En gran parte, los recientes avances en el área de aprendizaje profundo han permitido obtener resultados prometedores en esta línea de investigación. No obstante, los métodos del estado del arte se caracterizan por utilizar redes neuronales convolucionales muy complejas cuya etapa de inferencia no puede llevarse a cabo en tiempo real en sistemas con recursos computacionales limitados. Desde esta perspectiva, pocos trabajos se han propuesto para la estimación de profundidad monocular mediante arquitecturas de baja latencia. En el presente trabajo, se propone una arquitectura de red neuronal convolucional basada en el extractor de características ShuffleNet V2. Los resultados obtenidos en el conjunto de datos NYU Depth V2 muestran que la arquitectura propuesta es capaz de reducir el tiempo de inferencia en un 37.5% con respecto a métodos relacionados del estado del arte.

Palabras clave: Baja latencia, profundidad, monocular, ShuffleNet V2.

Low-latency Convolutional Neural Network for Monocular Depth Estimation

Abstract. Monocular depth estimation has recently become one of the emerging and increasingly interesting techniques for obtaining the spatial structure of a scene. To a large extent, recent advances in the area of deep learning have allowed promising results in this research area. However, state-of-the-art methods stand out for using highly complex convolutional neural networks whose inference stage cannot be deployed in real-time on systems with limited computational resources. From this perspective, few studies have been proposed to tackle the monocular depth estimation problem using low-latency architectures. In this paper, a convolutional neural network architecture based on the ShuffleNet V2 feature extractor is proposed. The results on the NYU Depth

V2 dataset show that the proposed architecture is able to reduce the inference time by 37.5% compared to related state-of-the-art methods.

Keywords: Low-latency, depth, monocular, ShuffleNet V2.

1. Introducción

La estimación de profundidad se considera como uno de los problemas fundamentales en el campo de la visión por computadora al tratarse de una función imprescindible para la realización de tareas tales como la localización, mapeo, detección de objetos 3D y aplicaciones de realidad aumentada [1]. En general, la estimación de profundidad se refiere al conjunto de técnicas y algoritmos diseñados con el propósito de obtener una representación de la estructura espacial de una escena. En otras palabras, el objetivo primordial de los métodos para la estimación de profundidad es obtener una representación de la distancia absoluta o relativa desde el sensor 3D a cada punto de la escena de interés [2].

La estimación de profundidad a partir de imágenes digitales se ha basado principalmente en un enfoque de visión estereoscópica, que requiere de un par de imágenes capturadas de la misma escena desde distintos ángulos a fin de triangular la posición 3D de cada pixel de las imágenes [2]. No obstante, en los últimos años, ha habido un creciente interés en la estimación de profundidad monocular relativa, es decir, utilizando una sola imagen RGB. Esto último se refiere a un problema inherentemente ambiguo, ya que no existe ninguna correlación entre la saturación de color y la profundidad relativa asociada a cada pixel en una imagen RGB [3].

Por un lado, el creciente interés en la estimación de profundidad monocular relativa se debe a la alta eficiencia energética, el tamaño reducido y el bajo costo de las cámaras monoculares. Por otro lado, los recientes avances en el campo del aprendizaje automático (en inglés, machine learning) han mostrado que es posible obtener mapas de profundidad a nivel de pixel a partir de una sola imagen, factor que también ha incrementado el interés por parte de la comunidad científica [1].

Desde la perspectiva del aprendizaje automático, se ha buscado resolver el problema de estimación de profundidad monocular por medio de técnicas tradicionales y de aprendizaje profundo (en inglés, deep learning). Los trabajos iniciales en esta línea de investigación propusieron métodos clásicos para la extracción manual de características de la imagen de entrada y modelos de Markov para la predicción del mapa de profundidad. Sin embargo, no fue hasta la implementación de técnicas de aprendizaje profundo cuando se lograron avances significativos [1]. Para tal efecto, las redes neuronales convolucionales (CNN por sus siglas en inglés) de tipo autoencoder se han convertido en el enfoque estándar para la resolución del problema en cuestión. En este sentido, el problema de estimación de profundidad monocular puede considerarse como un problema de regresión continua a nivel de pixeles que puede formularse de la siguiente manera [3]:

Sea I el espacio de imágenes RGB y D el codominio de mapas de profundidad de valor real. Dado un conjunto de datos de entrenamiento:

$$\mathcal{T} = \{(\mathbf{I}_i, \mathbf{D}_i)\}_{i=1}^{M}, \mathbf{I}_i \in I \quad y \quad \mathbf{D}_i \in D, \tag{1}$$

el objetivo es obtener la transformación no lineal dada por (2):

$$\varphi: I \rightarrow D \, . \tag{2}$$

Las redes neuronales convolucionales de tipo autoencoder que se utilizan actualmente para la estimación de profundidad monocular relativa están diseñadas para aproximar, de forma supervisada o no supervisada, la transformación no lineal descrita por la ecuación (2) [1]. La función de mapeo obtenida es capaz de reconstruir un mapa denso de profundidad relativa a partir de una sola imagen RGB. Las arquitecturas convolucionales de tipo autoencoder están conformadas por un codificador (en inglés, encoder) y un decodificador (en inglés, decoder). El codificador permite extraer los mapas de características que contienen la representación global de la imagen de entrada. Por su parte, el decodificador es la parte encargada de reconstruir un mapa denso de profundidad relativa a partir de la representación global extraída [4].

A pesar de que los resultados obtenidos han sido prometedores, los algoritmos del estado del arte están basados en arquitecturas de redes profundas muy complejas, cuya etapa de inferencia no puede llevarse a cabo en tiempo real en sistemas con recursos computacionales limitados. Este hecho impide su uso en campos de aplicación donde las características de las cámaras monoculares pudieran ser explotadas [1, 5, 6].

En este trabajo, se propone una arquitectura convolucional de tipo autoencoder de baja latencia que emplea el codificador ShuffleNet V2 [7] como el extractor de características. De acuerdo a la revisión de la literatura, este codificador originalmente diseñado para tareas de clasificación no ha sido evaluado en el problema de estimación de profundidad monocular relativa.

2. Trabajos relacionados

Saxena *et al.* [8] propuso el primer trabajo relacionado con la estimación de profundidad monocular mediante la aplicación de técnicas de extracción manual de características y un entrenamiento discriminativo de campos aleatorios de Markov. No obstante, el método en [8] se caracteriza por un bajo desempeño en ambientes no controlados. Por su parte, Karsch *et al.* [9] propuso un método basado en el algoritmo K-NN (K Nearest Neighbors) para la estimación de profundidad en imágenes con fondo estático. A pesar del avance que representaron dichos trabajos [8, 9], ambos requieren de procedimientos laboriosos de alineación.

No fue hasta la implementación de técnicas de aprendizaje profundo cuando se lograron avances significativos en esta línea de investigación. El primer trabajo que impulsó este enfoque [10], utilizó dos redes neuronales convolucionales acopladas para inferir la profundidad asociada a cada pixel en una imagen. Este trabajo sobresalió al ser el primero en introducir el concepto de información multiescala en la estimación de profundidad monocular.

Estudios posteriores, tal como [11], se enfocaron en obtener transiciones visuales más nítidas al incorporar campos aleatorios condicionales (en inglés, conditional random fields) como una etapa de regularización. Por su parte, Laina *et al.* [4] fue uno de los primeros grupos de investigación en proponer el uso de redes neuronales convolucionales de tipo autoencoder, las cuales se caracterizan por no requerir etapas de postprocesamiento u otros pasos de refinamiento como los métodos anteriormente

descritos. A partir de este punto, la arquitectura de red neuronal convolucional de tipo autoencoder se comenzó a establecer como uno de los métodos de mayor desempeño en esta línea de investigación.

Los trabajos posteriores al propuesto en [4] se han enfocado en mejorar la métrica de exactitud de diversas maneras, p. ej., a través de la inclusión de información semántica en la etapa de entrenamiento [12]. En particular, el trabajo de Fu *et al.* [13] fue bien recibido por parte de la comunidad científica al proponer una arquitectura robusta que, haciendo uso de una técnica de convolución dilatada, es capaz de proveer mapas de profundidad de alta resolución. De manera similar a [13], otros trabajos, tal como el de Kumari *et al.* [14], han explorado la optimización del modelo mediante funciones de pérdida perceptivas con el propósito de mejorar la retención de detalles locales. Común a los trabajos arriba mencionados, es la estrategia de aprendizaje supervisada, la cual requiere datos etiquetados para el entrenamiento de las redes neuronales convolucionales.

Por otro lado, la estrategia de aprendizaje no supervisada es el enfoque ortogonal al mencionado anteriormente. Este enfoque explota el potencial de las redes neuronales convolucionales sin requerir un conjunto de datos de entrenamiento etiquetados. En este sentido, los autores en [15] abordaron la estimación de profundidad monocular como un problema de reconstrucción de imágenes, para el cual definen una nueva función de pérdida capaz de imponer consistencia a las disparidades producidas entre las imágenes derecha e izquierda. Recientemente, investigadores de la Universidad de Oxford [16] propusieron un método novedoso para la estimación de profundidad monocular basado en una red generativa cíclica antagónica (CGAN o CycleGAN por sus siglas en inglés). Dicho método no supervisado, aprovecha la síntesis de imágenes para la medición indirecta de la disparidad.

Cabe destacar que los métodos previamente mencionados, se han enfocado en incrementar la métrica de exactitud de la etapa de validación a expensas de una elevada cantidad de operaciones computacionales [5]. En este sentido, un muy escaso número de estudios se han realizado con el objetivo de explorar arquitecturas de bajo costo computacional para la estimación de profundidad monocular relativa [1, 5, 6]. Por un lado, Poggy *et al.* [5] adopta una estrategia de aprendizaje no supervisada para entrenar una red neuronal convolucional compacta de tipo autoencoder con respecto al conjunto de datos KITTI [17], con la cual, logra una reducción significativa en el número de parámetros y con ello, un decremento en el tiempo de inferencia. De manera similar, Wofk *et al.* [6] propone una arquitectura de red neuronal convolucional de tipo autoencoder, cuya principal característica es el diseño de un decodificador de baja latencia. Asimismo, los autores en [6] sugieren el uso adicional de algoritmos de compresión para disminuir aún más el tiempo de inferencia.

Siguiendo un enfoque similar a [6], este artículo demuestra que el desempeño en el tiempo de inferencia puede ser reducido si se utiliza el codificador ShuffleNet V2 [7]. Este codificador incorpora canales barajados (en inglés, channel shuffle) originalmente diseñados para reducir el costo computacional en tareas de clasificación. En general, la arquitectura convolucional de tipo autoencoder propuesta logra un tiempo de inferencia menor al que se logra con la arquitectura propuesta en [6]. Además, los resultados obtenidos demuestran que es factible obtener un tiempo de inferencia incluso menor a aquel obtenido mediante el uso adicional de algoritmos de compresión.

3. Metodología

En esta sección, se describe la arquitectura de la red neuronal convolucional propuesta, el conjunto de datos de entrenamiento seleccionado, el protocolo de entrenamiento establecido y las métricas de evaluación.

3.1. Arquitectura de la red

Las arquitecturas convolucionales de tipo autoencoder que se utilizan para la tarea en cuestión están conformadas de una etapa de codificación seguida de una etapa de decodificación [4]. La etapa de codificación permite extraer los mapas de características que contienen la representación global de la imagen RGB de entrada (altura x anchura x 3). Estos mapas de características son alimentados a la etapa de decodificación para reconstruir un mapa denso de profundidad (altura x anchura x 1) mediante operaciones de convolución e interpolación. La arquitectura propuesta se ilustra en la Fig. 1.

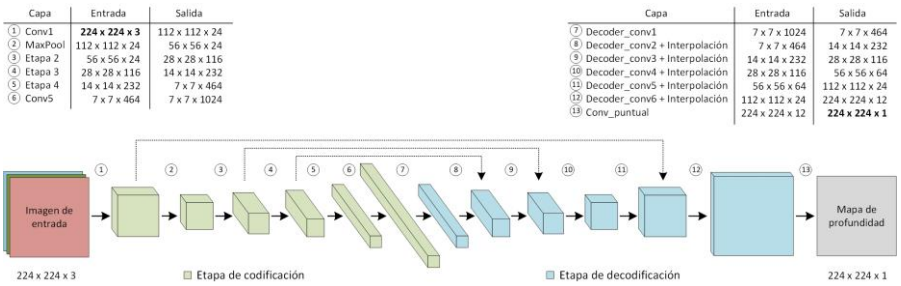


Fig. 1. Propuesta de arquitectura CNN de tipo autoencoder.

Para la etapa de extracción de características, se seleccionó el codificador del estado del arte ShuffleNet V2 [7]. Según los autores en [7], el codificador fue diseñado siguiendo cuatro pautas generales que se listan a continuación:

1. La profundidad/anchura de los canales en ambos extremos de una capa debe mantenerse igual para minimizar el costo de acceso a la memoria (MAC).
2. Se debe evitar el uso excesivo de grupos de convolución ya que esto aumenta el costo de acceso a la memoria.
3. Se sugiere minimizar el uso de operadores fragmentados para mantener un alto grado de paralelismo.
4. Se recomienda reducir el uso de operaciones elemento a elemento ya que su costo computacional no es despreciable.

En general, los autores en [7] sugieren utilizar métricas directas (p. ej., MAC) en lugar de métricas indirectas como el número de operaciones de coma flotante por segundo (FLOPS por sus siglas en inglés) para el diseño de arquitecturas de baja latencia. Nótese que la última capa de propagación hacia delante (en inglés, fully connected) del codificador ShuffleNet V2 debe ser removida para un correcto acoplamiento con la etapa de decodificación.

Además, los autores en [7] mencionan que es posible configurar un factor de profundidad que ajusta el número de canales de cada bloque. En este caso, el factor se estableció en *1x* con el propósito de mantener el número de parámetros de la red en un valor relativamente bajo y de esta manera, minimizar el tiempo de inferencia sin comprometer el valor de la exactitud umbral.

Por otro lado, la etapa de decodificación (ver Fig. 1) está inspirado en el diseño propuesto en [6]. En contraste con este último, el decodificador utilizado está conformado por seis capas de operaciones de convolución e interpolación. Estas capas están encargadas de aumentar gradualmente la resolución y reducir el número de canales. Cada una de estas capas utiliza un filtro de convolución de 5x5 y lleva a cabo la interpolación mediante el método del vecino más cercano. Un filtro de convolución puntual de 1x1 es aplicado al último mapa de características (224 x 224 x 12) para obtener el mapa de profundidad relativa resultante.

Similar a [6], se agregaron tres conexiones residuales aditivas entre el codificador y el decodificador que se ilustran como flechas punteadas en la Fig. 1. Estas conexiones permiten contrarrestar el problema de fuga de gradientes y la consecuente pérdida de nitidez en el mapa de profundidad relativa [14, 6].

3.2. Conjunto de datos de entrenamiento

El conjunto de datos seleccionado para entrenar y evaluar la arquitectura convolucional de tipo autoencoder propuesta es la NYU Depth V2 [18]. Se trata de un conjunto de datos disponible al público a partir del año 2012 y desde entonces, ha sido ampliamente usado en los trabajos del estado del arte relacionados con la estimación de profundidad monocular supervisada [18]. El conjunto de datos NYU Depth V2 está conformado por secuencias de video de diversas escenas de interiores que han sido capturadas mediante una cámara RGB-D. El conjunto de datos oficial contiene 120,000 muestras de entrenamiento y 654 muestras de validación, donde cada muestra consiste en un par de imágenes alineadas RGB y de profundidad. No obstante, algunos trabajos del estado del arte utilizan un subconjunto de 50,000 muestras de entrenamiento seleccionadas aleatoriamente para reducir el tiempo de dicha etapa [19]. En ese sentido, el método propuesto es entrenado bajo dicho esquema.

3.3. Protocolo de entrenamiento

La red neuronal convolucional propuesta es entrenada bajo una estrategia de aprendizaje supervisada. En particular, el proceso de entrenamiento fue llevado a cabo mediante el uso de una laptop Asus X556U 8GB RAM equipada con una tarjeta de video dedicada NVIDIA GeForce 930MX con una VRAM de 2GB. La implementación de la arquitectura propuesta se realizó mediante PyTorch 1.4.0 [20] (CUDA habilitado) en el entorno de desarrollo Spyder 4 y la distribución Ubuntu 16.04 LTS de Linux.

A partir de la interfaz de programación que provee Pytorch, se seleccionó el método del descenso de gradiente estocástico como el optimizador para la etapa de entrenamiento. Como parte de los hiperparámetros del optimizador, el impulso fue definido en 0.95, mientras que el método de regularización L2 fue habilitado con un valor de 0.0001.

Por las limitaciones propias de la VRAM descrita anteriormente, el tamaño del lote para la etapa de entrenamiento (en inglés, *batch size*) fue establecido en 8. Por su parte, el número máximo de épocas se definido en 30. Finalmente, la tasa de aprendizaje se estableció en 0.01 y fue configurada para que disminuyera su valor en un factor de 10 cada 5 épocas transcurridas. Los valores de los hiperparámetros anteriormente descritos fueron obtenidos a través de pruebas experimentales. Similar a otros trabajos del estado del arte [19], los pesos del codificador son inicializados a partir de un modelo preentrenado en el conjunto de datos ImageNet [21].

El error absoluto medio (MAE por sus siglas en inglés) se seleccionó como la función de pérdida a minimizar durante la etapa de entrenamiento [22]. Esta función considera la diferencia entre el mapa de profundidad objetivo que representa la verdad fundamental y (en inglés, ground truth) y la predicción de la red $\hat{y}$. La función de pérdida está dada por la siguiente expresión:

$$L_1(y, \hat{y}) = \frac{1}{n}\sum_{p=1}^{n}\left|y_p - \hat{y}_p\right|, \tag{3}$$

donde y_p es un pixel del mapa de profundidad objetivo y, $\hat{y}_p$ es un pixel en el mapa de profundidad relativa estimado $\hat{y}$, y n es el número total de pixeles del mapa de profundidad relativa.

3.4. Métricas de evaluación

El método propuesto se compara contra [6] mediante el tiempo de inferencia en milisegundos (ms) y dos métricas estándar en la estimación de profundidad monocular que se definen a continuación:

— Raíz del error cuadrático medio (RMSE):

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{p=1}^{n}\left(y_p - \hat{y}_p\right)^2} \tag{4}$$

— Exactitud umbral (δ_1):

$$\%\ de\ y_p \mid \max\left(\frac{y_p}{\hat{y}_p}, \frac{\hat{y}_p}{y_p}\right) = \delta_1 < 1.25, \tag{5}$$

donde y_p es un pixel del mapa de profundidad objetivo y, $\hat{y}_p$ es un pixel en el mapa de profundidad relativa estimado $\hat{y}$, y n es el número total de pixeles del mapa de profundidad relativa.

La exactitud umbral debe entenderse como el porcentaje de pixeles del mapa de profundidad estimado para los cuales el error relativo es menor o igual al 25% [1, 6].

4. Resultados

En esta sección, se presentan los resultados cuantitativos y cualitativos que se obtuvieron al evaluar el modelo obtenido de la etapa de entrenamiento contra el

conjunto de datos de validación. Los resultados se contrastan con [6] al ser el único estudio relacionado que entrena y evalúa una arquitectura de baja latencia con respecto al conjunto de datos NYU Depth V2.

El proceso de evaluación se llevó a cabo utilizando la misma configuración de hardware que se utilizó para la etapa de entrenamiento.

3.1. Resultados cuantitativos

La Tabla 1 muestra la comparativa entre los resultados cuantitativos obtenidos al evaluar el método relacionado del estado del arte y el método propuesto contra el conjunto de datos de validación que proporciona la base de datos NYU Depth V2. Similar a los trabajos del estado del arte, cada valor en la Tabla 1 representa el promedio obtenido en el conjunto de datos de validación.

Tabla 1. Comparativa de resultados cuantitativos en el conjunto de datos de validación.

Método	$\boldsymbol{\delta_1} \uparrow$	**RMSE** ↓	**GPU [ms]** ↓	**fps** ↑
Wofk *et al.* [6]	0.775	591.619	24	41
Wofk *et al.* + PN [6]	0.771	603.741	17	58
Método propuesto	0.757	606.351	15	66

El valor de la exactitud umbral (δ_1) obtenido por el método propuesto resulta ligeramente menor al valor alcanzado por el método del estado del arte (sin compresión). En ese sentido, el valor de la raíz del error cuadrático medio (RMSE) resulta ligeramente menor a favor del método del estado del arte (sin compresión). Dado que la magnitud de las diferencias porcentuales para cada métrica mencionada no es significativa, estas podrían considerarse despreciables.

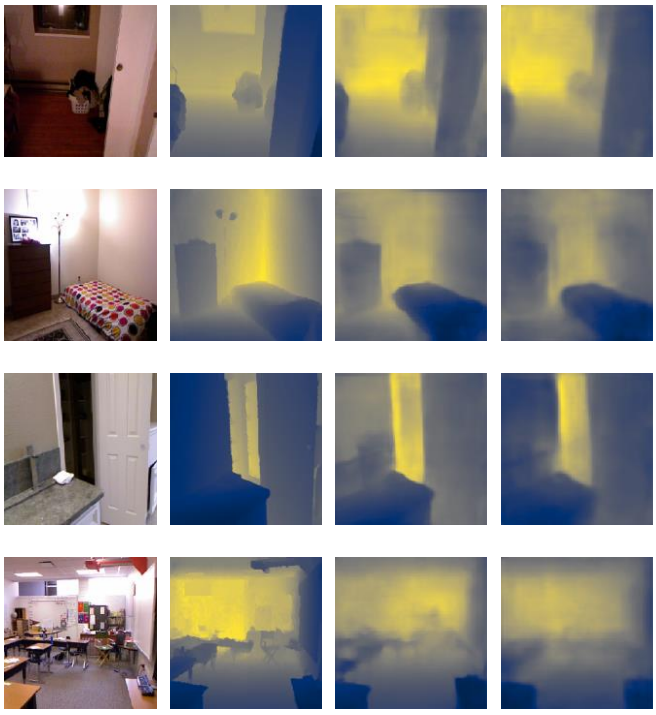
Por su parte, el tiempo de inferencia (ms) logrado por el método propuesto es menor al obtenido por el método del estado del arte (sin compresión) en un 37.5%. Nótese que el tiempo de inferencia del método propuesto es incluso menor al que se obtiene utilizando el modelo comprimido (poda neuronal) del método del estado del arte en un 11.7%.

3.2. Resultados cualitativos

La Fig. 2 muestra una comparativa de los resultados cualitativos con el método relacionado del estado del arte. Debe notarse que la profundidad relativa en cada escena disminuye conforme el color de los pixeles en los mapas de profundidad se oscurece. En otras palabras, entre más oscuro sea el color del pixel, más cercanos están los objetos en la escena.

En general, se aprecia en la Fig. 2 gran similitud entre el mapa de profundidad correspondiente al método propuesto (d) con respecto al método del estado del arte (c), aunque con mínimas diferencias en ciertas regiones locales.

Debe destacarse que ninguna imagen de la Fig. 2 fue utilizada en la etapa de entrenamiento.



(a) (b) (c) (d)

Fig. 2. Comparativa de resultados cualitativos en el conjunto de datos de validación (NYU Depth V2). De izquierda a derecha: (a) imagen RGB de entrada, (b) mapa de profundidad objetivo, (c) Wofk *et al.*+PN [6] y (d) método propuesto.

Desde esta perspectiva, es importante resaltar la buena capacidad de generalización del modelo obtenido en escenas de interiores del conjunto de datos de validación. Nótese que este modelo podría no funcionar de manera adecuada en escenas de exteriores.

5. Conclusiones y trabajo a futuro

El presente trabajo propone una arquitectura de red neuronal convolucional de tipo autoencoder para la estimación de profundidad monocular relativa en sistemas con recursos computacionales limitados. Los resultados experimentales del enfoque propuesto muestran que los filtros de convolución de baja latencia que conforman el codificador ShuffleNet V2 pueden contribuir a sentar las bases para el diseño de arquitecturas convolucionales para la estimación de profundidad monocular relativa en tiempo real. Nótese que el codificador ShuffleNet V2 no había sido evaluado en el problema de estimación de profundidad monocular, hasta ahora.

Debe destacarse que, sin el uso adicional de técnicas de compresión, el modelo obtenido se caracteriza por un tiempo de inferencia menor al que se obtiene al evaluar el método relacionado del estado del arte, manteniendo una exactitud umbral comparable al de dicho trabajo. En particular, el método propuesto logra una reducción del 37.5% en tiempo de inferencia con respecto al método mencionado anteriormente. Asimismo, debe notarse que el tamaño reducido de la arquitectura propuesta permite llevar a cabo las etapas de entrenamiento y validación en una laptop con especificaciones técnicas muy por debajo a las requeridas por los métodos del estado del arte que no son considerados de baja latencia.

Como trabajo a futuro, se tiene considerado entrenar la arquitectura propuesta en el conjunto de datos de entrenamiento KITTI [17] para evaluar su desempeño en escenas de exteriores. Asimismo, se pretende mejorar el desempeño de la métrica de exactitud umbral mediante la implementación de un decodificador que sea diseñado bajo las pautas y lineamientos sugeridos por los autores en [7].

Agradecimientos. Los autores agradecen a la Universidad Autónoma de Querétaro y al Consejo Nacional de Ciencia y Tecnología (CONACYT) por el apoyo recibido durante el desarrollo de este proyecto de investigación. (Beca CONACYT No. 2019-000037-02NACF-03968).

Referencias

1. Zhao, C., Sun, Q., Zhang, Tang, Y., Qian, F.: Monocular depth estimation based on deep learning: an overview (2020)
2. Giancola, S., Valenti, M., Sala, R.: A survey on 3d cameras: metrological comparison of time-of-flight, structured-light and active stereoscopy technologies. Springer, Cham (2018)
3. Bhoi, A.: Monocular depth estimation: a survey (2019)
4. Laina, I., Rupprecht, C., Belagiannis, V., Tombari, F., Navab, N.: Deeper depth prediction with fully convolutional residual networks. In: Proceedings of the IEEE International Conference on 3D Vision (3DV), pp. 239–248 (2016)
5. Poggi, M., Aleotti., F., Tosi, F., Mattoccia, S.: Towards real-time unsupervised monocular depth estimation on CPU. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 5848–5854 (2018)
6. Wofk, D., Ma, F., Yang, T., Karaman, S., Sze, V.: FastDepth: fast monocular depth estimation on embedded systems. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), pp. 6101–6108 (2019)
7. Ma, N., Zhang, X., Zheng, H.T., Sun J.: ShuffleNet V2: Practical guidelines for efficient cnn architecture design. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds) Computer Vision – (ECCV) Lecture Notes in Computer Science, 11218, Springer, Cham (2018)
8. Saxena, A., Chung, S.H., Ng, A.Y.: Learning depth from single monocular images. In: Advances in neural information processing systems 19 (NeurIPS), pp. 1161–1168, Curran Associates, Inc. (2006)
9. Karsch, K., Liu, C., Kang, S.B.: Depth extraction from video using non-parametric sampling. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds) Computer Vision – (ECCV) Lecture Notes in Computer Science, 7576, Springer, Berlin, Heidelberg (2012)

10. 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 27 (NeurIPS), pp. 2366–2374, Curran Associates, Inc. (2014)
11. 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)
12. Gurram, A., Urfalioglu, O., Halfaoui, I., Bouzaraa, F., López, A. M.: Monocular Depth Estimation by Learning from Heterogeneous Datasets. In: Proceedings of the IEEE Intelligent Vehicles Symposium (IV), pp. 2176-2181, IEEE (2018)
13. Fu, H., Gong, M., Wang, C., Batmanghelich, K., Tao, D.: Deep Ordinal Regression Network for Monocular Depth Estimation. In: Proceedings of the IEEE/CFV Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2002-2011, IEEE (2018)
14. Kumari, S., Jha, R. R., Bhavsar A., Nigam, A.: AUTODEPTH: Single Image Depth Map Estimation via Residual CNN Encoder-Decoder and Stacked Hourglass. In: Proceedings of the 26th IEEE International Conference on Image Processing (ICIP), pp. 340-344, IEEE (2019)
15. Godard, C., Aodha, O. M., Brostow, G. J.: Unsupervised Monocular Depth Estimation with Left-Right Consistency. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6602-6611, IEEE (2017)
16. Pilzer, A., Xu, D., Puscas, M., Ricci, E., Sebe, N.: Unsupervised Adversarial Depth Estimation Using Cycled Generative Networks. In: Proceedings of the IEEE International Conference on 3D Vision (3DV), pp. 587-595, IEEE (2018)
17. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: The KITTI dataset. International Journal Robotics Research, 32(11), 1231–1237 (2013)
18. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor Segmentation and Support Inference from RGBD Images. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds) Computer Vision – ECCV 2012. ECCV 2012. Lecture Notes in Computer Science, vol. 7576. Springer, Berlin, Heidelberg (2012)
19. Alhashim, I., Wonka, P.: High Quality Monocular Depth Estimation via Transfer Learning. arXiv preprint arXiv: 1812.11941v2 (2018)
20. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: PyTorch: An Imperative Style, High-Performance Deep Learning Library. In: Advances in Neural Information Processing Systems 32 (NeurIPS), pp. 8024-8035, Curran Associates, Inc. (2019)
21. Deng, J., Dong, W., Socher, R., Li, L., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 248-255, IEEE (2009)
22. Carvalho, M., Saux, B. L., Trouvé-Peloux, P., Almansa, A., Champagnat, F.: On Regression Losses for Deep Depth Estimation. In: Proceedings of the 25th IEEE International Conference on Image Processing (ICIP), pp. 2915-2919, IEEE (2018)