Sensors & Transducers, Vol. 263, Issue 4, December 2023, pp. 105-118

Sensors & Transducers

Published by IFSA Publishing, S. L., 2023
http://www.sensorsportal.com

Double Input Convolutional Neural Network Model for Determining Heights of Figures in 2D Images from their Shadows

[1, *] Julián René MUÑOZ BURBANO, [2] Pablo Emilio JOJOA GÓMEZ and Fausto Miguel CASTRO CAICEDO

[1] Corporación Universitaria Comfacauca Unicomfacauca, Group MIND, Centro Histórico. Popayán, Colombia
[2] Universidad del Cauca, Telecommunications Group I+D-GNTT, Campus Universitario Tulcán. Popayán, Colombia
[1] Tel.: (57)3108987728
E-mail: jburbano@unicomfacauca.edu.co

Received: 2 October 2023 / Accepted: 7 November 2023 / Published: 21 December 2023

Abstract: This research developed a model using convolutional neural network (CNN) architecture for pattern detection and feature extraction from shadows and shapes in images. The dataset was constructed with photographic images of figures, with their respective shadows cast from different angles and locations. A DataFrame was constructed that combined with images of the figures improves the accuracy of the algorithm. Therefore, the presented model offers a favorable structure for estimating the height of 2D figures from shadows in images. The convolutional design makes it suitable for learning complex patterns in visual data, while the linear activation function in the output layer enables regression tasks. Thus, the optimized CNN model represents an automated process that learns to map patterns in shadow images to corresponding heights

Keywords: Shadows, Shapes, Convolutional neural network, Dataset, DataFrame, 2D, Height.

1. Introduction

When estimating the height of a figure or object contained in a 2D image from its shadow is a complex process and is encountered under different lighting conditions. This has become a major challenge, as the need for an accurate model that can estimate the geometry of the object and the way the light hits it. However, estimating the height of a figure or object in a 2D image from its shadow becomes a more complex problem due to the lack of information. In this context, the use of neural networks has become a valuable tool to address this research problem, and a solution based on convolutional neural networks for height prediction from its shadow is presented. The proposed model uses a convolutional neural network architecture with regression to extract relevant image and shadow features to perform figure or shape height prediction.

In [31] (Khan, M.S.U. et al. 2022) Three-dimensional (3D) reconstruction from a two-dimensional (2D) image is a prominent challenge in the field of image processing and computer vision. Traditionally considered an ill-conditioned problem in terms of optimization, recent advances in machine learning techniques have significantly boosted the efficiency of 3D reconstruction. The complexity rises when dealing with objects with complex deformations or lacking texture, due to the possibility of projecting

multiple 3D objects from the same 2D plane. This research provides a comprehensive review of studies conducted between 2018 and 2021 related to 3D reconstruction from a single perspective, especially highlighting deep learning-based methods. A barrier to direct comparison between the reviewed methods lies in the absence of standardized datasets and the diversity of methods for representing 3D shapes.

The review covers different reconstruction techniques, such as depth map generation, surface normal calculation, point cloud creation and mesh shaping. It also examines various loss functions and metrics that are employed for both training and evaluation of deep learning models in the 3D reconstruction task. This analysis provides a detailed overview of current approaches and their methodologies, highlighting the evolution and progress in this field of research. It also formulates several interesting conclusions, such as three-dimensional (3D) reconstruction from a monocular image is recognized as a complex problem in the field of computer vision, largely due to the lack of a standardized approach to represent 3D shapes and the scarcity of datasets covering a sufficient diversity of objects. This paper argues that real-world 3D data are limited and collecting them is inherently challenging, especially in the case of untextured surfaces, where datasets are even scarcer. There are also generative antagonistic networks (GANs), which have proven to be remarkably efficient and fast, reaching point cloud generation frequencies of 250 Hz, which is favorable for real-time applications. In addition, the adoption of transformers in vision tasks is emerging as a promising technique, with a specific network improving low-resolution 3D reconstruction and intersection index over union (IoU) on the ShapeNet dataset. In summary, it is a paper that suggests that there is a promising path towards improved monocular 3D reconstruction, although there remains a need for expansion of datasets and standardization of techniques to effectively compare methods and results between different investigations.

For the training and validation of the proposed model, a dataset was used that includes photographs of a diversity of geometric figures, multiple shapes and shadows under various lighting conditions. With this dataset, the validation training and testing process of the model was performed to have the ability to estimate the height of the figures, which suggests or represents a promising solution for estimating the height of figures in two-dimensional images based on their shadows. Future research will focus on judging the accuracy and efficiency of this innovative method by comparing it to other pre-existing algorithms that also estimate the height of objects from their shadows.

The proposed model was evaluated with a test data set, achieving remarkable accuracy in estimating object heights compared to previous methods. In essence, the integration of cast shadow data in a learning model plus a DataFrame represented a very good strategy when estimating the height of a figure from its shadow in 2D images.

By applying advanced deep learning techniques, it is possible to process and analyze large volumes of data, leading to increased accuracy in estimating the height of shapes from their shadows. Furthermore, the ability to integrate new geometric shapes and shadows within the image set and used in the model provides an additional variety of information for future research, especially when working with irregular shapes.

Similarly, in [9] (S. Mohajerani, P. Saeedi. 2018), the automatic detection of shadows in images using a segmentation method and the application of Deep Learning is addressed. The study focuses on identifying pixel-level shadow regions in RGB images by extracting shadow features and providing a Convolutional Neural Network (CNN) with the ability to detect shadow patterns both locally and globally.

On the other hand, in [25] (Tao, M. W., Srinivasan, P. P., Hadap, et al. 2017), they developed an algorithm based on dense depth estimation which combines blur and correspondence metrics. In addition, it defines an optimization framework that integrates photographic, depth and shading coherence for light depth-of-field estimation in different scenarios.

1.1. The Shadow

Shadows are defined as segments of a scene that are not directly illuminated by a light source. This definition is extended to identify two distinct categories of shadows. The first is the intrinsic shadow, characterized by the fact that the shadow area is located on the object itself that generates the shadow. The second is cast shadow, which refers to the area of shadow visible on surfaces other than the obstructing object, either the background or other surrounding objects. Within the cast shadow, it is common to differentiate two areas: the umbra, which is the darkest part of the shadow where light, is completely blocked, and the penumbra, which is the transition zone where Plight, is partially blocked (Fig. 1).

Fig. 1. Figure and corresponding shadow and Rotation of the light source.

In [4] (Vicente and Samaras, 2014)], recognizes that two-dimensional digital images have features such as textures, edges, shapes, colors and shadows. The latter, in particular, provide critical data that can be exploited by shadow detection techniques to infer relationships between object shape, light sources, and shadowed areas.

Similarly, previous research such as those of [8] (Kriegman and Belhumeur. 1998) and [11] (Knill et al. 1997), cited respectively, have given importance of shadows as indicators of information about the shape and topography of surfaces and objects. These help to identify areas of interest, establish the direction of light sources, and discern geometric details. In this context, the work of (Knill, Mamassian and Kersten. 1997) and Shafer and Kanade (1983), [11] (D. C. Knill, et al. 1997) and [17] (S. A. Shafer, T. Kanade. 1983), respectively, have shown that the intrinsic geometric properties of shadows are particularly valuable. This is because they facilitate the understanding of the interaction between shapes, shadows and illumination, contributing to the perception of the structure and height of the surfaces under study.

In numerous previous works, as indicated in [24] (Hintze, Morse. 2019), the information extracted from the shadow using techniques, methods and algorithms for purposes such as: determining the location of the shape or object, area of interest, direction of the light source, among others, is usually oriented towards its subsequent elimination, where this makes its significance not so remarkable.

In [30] (Huang, X. et al. 2007) presents an approach for shape-from-shading (SFS) estimation. They propose an example-based method that increases the robustness and accuracy of SFS. The heart of this method lies in the use of a database containing synthetically generated appearances of known three-dimensional models. Each pixel in these synthetic images is linked to its corresponding real surface normal. By confronting the input image with this database, the algorithm is able to identify the normal for each point in the image. The method has been tested using a small generic database consisting of 50 spheres of different radii. The results have shown a significant improvement in the quality of reconstruction of both synoptic and real images. The method is particularly effective in reconstructing shapes with minute details, validating its potential for applications in environments where traditional FSS methods fail. The presented study introduces an advanced Shape-from-Shading (SFS) algorithm that improves the estimation of surface gradient maps for intensity-based images. The technique employs sample images with known surface normal as an additional source of information, incorporating them as a constraint term in an energy-minimizing process.

On the other hand, in [12] (Salvador, Cavallaro and Ebrahimi. 2004) it is stated that the shadows in a 2D image contain relevant information about the scene, obtaining the location of the shape or object, the characteristics of the surface and the light source. Techniques are mentioned that are based on models that represent the knowledge of the scene geometry, objects or shapes, the light source, as well as techniques based on properties that identify shadows using characteristics such as geometry, brightness and color of the shadows (Fig. 2).

Similarly, in [12] (Salvador, Cavallaro and Ebrahimi. 2004), general aspects related to digital image processing are addressed, using convolutional neural networks (CNN) with different data sets (datasets), where the latter are mainly characterized by containing images of geometric primitives and the shadows, they cast according to an illumination source located in a specific location. Similarly, in [2] (Panagopoulos, Hadap, Samaras and Dimitris. 2013) in their research they use synthetic images divided into small rectangular regions, and thus, for each of these it is possible to capture what corresponds to the distribution of intensities and the ability to handle surfaces. At this point, the geometry sections are combined, integrating the intensity distribution into a dictionary that, with all the sections, can generate a hypothesis about that shape.

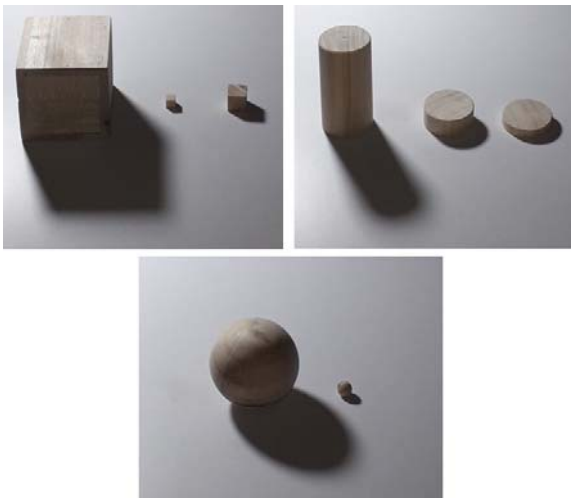

Fig. 2. Types of dataset figures.

Employing the learning model that processes the images of the training set, which with all sections or small regions form a large dictionary, and where the decomposition of an image into shadow and reflectance images describe the brightness perception for these images.

In [1, 13] (Hosseinzadeh, S., Shakeri, M. and Zhang, H. 2005) it is shown that the concept of sections corresponds to a collage of sections with a different reflectance and which are illuminated with an illumination source that varies slowly in intensity, and with shapes that are of a constant scale factor. This learning model with the data set to be generated would be used to determine the surface height and 3D generation of the object or shape, initially taking the information of the shadow it casts on a 2D image. In [12] (Salvador, Cavallaro, Ebrahimi, Touradj. 2004), [7] (G. Liasis, S. Stavrou. 2016) illumination, shadows and reflectance are an integral part of the different scenes, providing relevant information about the shape and appearance of the object.

An innovative deep learning architecture designed to transform a color image into a 3D shape composed of a triangular mesh is presented in [33]. Previous approaches in this field were often limited to 3D shape

representations such as volumes or point clouds, which do not translate easily into usable mesh models. The architecture proposed by the authors departs from these limitations by implementing a graph-based convolutional neural network that generates 3D meshes. This method innovates by progressively deforming an ellipsoid using detailed perceptual features obtained from the input image. To ensure the stability of the deformation process, the researchers adopt a strategy that goes from general to specific (coarse to fine). In addition, various loss functions linked to the mesh structure are introduced, which aim to ensure that the resulting 3D geometry is not only visually appealing, but also physically accurate. Extensive experiments demonstrate that the proposed method is capable of producing mesh models with a superior level of detail and accuracy in 3D shape estimation that outperforms pre-existing state-of-the-art techniques.

In [2] (Panagopoulos, Hadap, and Samaras. 2013) propose the use of a dictionary use of a dictionary (Dictionary) that is confirmed by different patches (Patches) of geometry associated to the image, to a distribution of pixel intensities that can generate that geometry and with the use of larger image regions that allow capturing appearance data, it is possible with a data set and supported in neural networks it is possible to design a model that fits the proposed research.

In [22] (Varol, Shaji, Salzmann, Fua. 2011) has an approach based on deformable shape recovery that can work under complex illumination and on partially textured surfaces. They used an algorithm that performs a learned mapping of the intensity distribution and shape of local surface patches, which is focused on 3D surface shape estimation.

In [14] (D. S. Kim, M. Arsalan, K. R. Park. 2018) they employ a Convolutional Neural Network (CNN) for shadow detection using cameras and using the open database CAVIAR (Context-Aware Vision Using Image-Based Active Recognition). They indicate that in their research it is not computationally expensive both in training and validation and can be adapted to tasks such as image segmentation and has a level of accuracy of other methods.

In [34] (Anny Yuniarti, Nanik Suciati 2019) is a review of 3D reconstruction. Although deep learning techniques have revolutionized image segmentation and object recognition, their application in 3D reconstruction is at a nascent stage. Nevertheless, their potential is considerable and their exploration promises to significantly improve the performance of reconstruction methods. [34] focuses on the literature review specializing in 3D reconstruction from single or multiple images, explicitly excluding RGB-D input. This article is distinguished as one of the first systematic reviews focused on 3D reconstruction through deep learning, providing a comprehensive overview of the most recent and advanced work in this field. The conclusion emphasizes that, although remarkable progress has already been made with the use of deep learning techniques in 3D reconstruction, there is room for improvement.

The content presents an overview of the evaluation metrics used to measure the effectiveness of different representations and methods in 3D reconstruction, focusing on three main types: point clouds, voxels and meshes. Voxel representation faces limitations in resolution and memory consumption, problems that can be mitigated by using octrees or learning to encode and decode 3D point representations, although the latter lacks surface connectivity. On the other hand, meshes have surface connectivity, but involve a combinatorial optimization challenge due to the sampling of vertices from the object surface. Comparing different approaches, it is concluded that ResMeshNet outperforms both the point cloud-based approach (e.g., PSG) and other mesh-based methods (e.g., AtlasNet).

1.2. The Dataset

The experimentation process is carried out under controlled conditions in terms of illumination, shapes and surface. Such conditions were used in the model allowing initially to test the model of the network and that it learned certain characteristics, which is why it was decided to use the different photographs of three geometric shapes on a uniform surface, with a dimmable LED lamp and the shadows corresponding to each shape (Fig. 3). It is important to mention that the data set was oriented with the objective of capturing photographs under different lighting conditions to give an initial knowledge to the neural network.

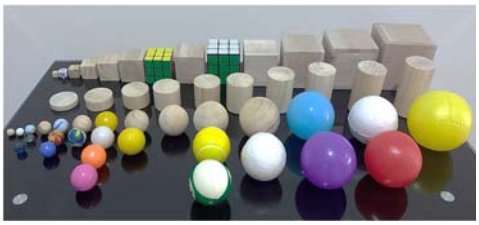

Fig. 3. Characteristics of the Dataset.

The data set includes 800×600 color photographs (RBG) of various geometric shapes of various colors, taken from a distance of 90 cm from the edge of the surface. The surface where the shape or object is located is white, smooth, does not reflect light, and is made of white matte adhesive vinyl. The surface has a size of 90 cm and a diameter of 180 cm, was designed circular to facilitate the displacement of the LED lamp, and is made of MDF (Medium Density Board) wood of 9 mm (Fig. 4).

The model integrates 16 layers, has a total number of parameters: 153985, all parameters are trainable. The architecture is designed to extract and learn features from both the images of the figures and their respective shadows, then combining this information to make a prediction of the height of the figures (see

Fig. 5). In [28] (René, J. et al. 2023) this model has a dual-input CNN is structured to perform regression tasks on images. It was designed with the extraction of significant features from two different input sources in mind, processing them in parallel and then combining this extracted information to generate a unified and better-informed prediction. With nearly 154000 trainable parameters, this architecture offers deep learning capability while maintaining a manageable, efficient size for training and deployment. Its modular and structured design facilitates fine-tuning, as well as additional optimizations to further improve performance and accuracy for specific image regression tasks.

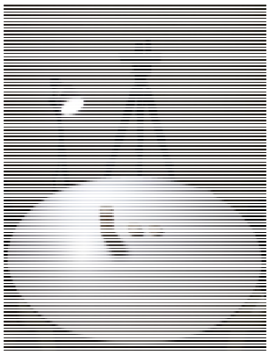

Fig. 4. Mock-up for taking photographs.

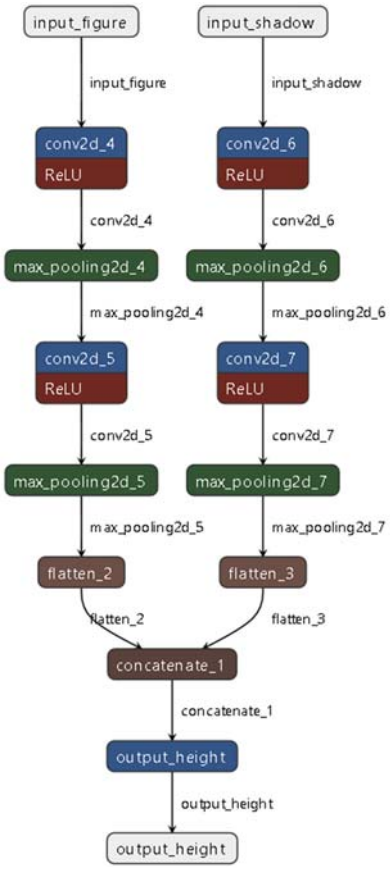

Fig. 5. Dual-Input Convolutional Neural Network Architecture.

1.3. Model Description

With the research conducted, it is proposed to use and/or design a convolutional neural network, which through the training of a set of data will provide the necessary knowledge to predict the height of an object or shape with the projected shadow. With the learning process, it is expected that the network, by using as input information the shadow cast by the shape or object in a 2D image, will predict or determine the height of the object in question. The data set (Dataset) was designed and developed with objects and/or basic geometric shapes (sphere, cube, cylinder) to perform the experimentation in a controlled manner, since after this proposal and as a future work it is expected to increase the data set with more complex shapes and shadows. The total amount of images in the dataset is approximately 4500 color photographs and for each shape, there are 1500 photographs (see Table 1), [28] (René, J. et al. 2023).

Table 1. Height Distribution of Geometric Shapes.

Figure	Color	Measurement (cm)	Quantity
Cube	Brown	1	1
Cube	Brown	1,5	1
Cube	Brown	2	1
Cube	Brown	3	1
Cube	Brown	4	1
Cube	Brown	5	1
Cube	Rubik's Cube	5,5	1
Cube	Brown	6	1
Cube	Rubik's Cube	6,5	1
Cube	Brown	7	1
Cube	Brown	8	1
Cube	Brown	9	1
Cube	Brown	10	1
Total Cubes			**13**
Cylinder	Brown	1	1
Cylinder	Brown	2	1
Cylinder	Brown	3	1
Cylinder	Brown	4	1
Cylinder	Brown	5	1
Cylinder	Brown	6	1
Cylinder	Brown	7	1
Cylinder	Brown	8	1
Cylinder	Brown	9	1
Cylinder	Brown	10	1
Cylinder	Brown	10,5	1
Total Cylinder			**11**
Sphere	Brown	1	1
Sphere	White	1,5	1
Sphere	White	2	3
Sphere	Brown	2,5	1
Sphere	White, Yellow, Blue	3	3
Sphere	Brown, Blue with yellow	3,5	2
Sphere	Yellow, White, Orange, Pink	4	4
Sphere	Brown, Yellow	4,5	2
Sphere	Brown	5	1
Sphere	Brown, Yellow, Green, White	6	3
Sphere	Brown, White	7	2
Sphere	Blue, Violet	8	2
Sphere	White, Red	8,5	2
Sphere	Yellow	9	1
Total Sphere			**28**

It is important to note that the proposed model (Figs. 6 and 7) uses geometry of three-dimensional shapes of the objects or shapes with the corresponding shadow from different locations in the plane and generated by a single light source. The latter is rotated around the figure to obtain shadows from different angles. The light source is located 90 centimeters from the center of the object to the base of the lamp and the heights of the lamp vary from 45 centimeters (cm) to 50 centimeters (cm) so that the different shadows are observable on the surface under varying intensities of illumination (see Fig. 1), [28] (René, J. et al. 2023).

Layer (type)	Output Shape	Param #	Connected to
input_figure (InputLayer)	[(None, 128, 128, 3)	0	
input_shadow (InputLayer)	[(None, 128, 128, 3)	0	
conv2d (Conv2D)	(None, 126, 126, 32)	896	input_figure[0][0]
conv2d_2 (Conv2D)	(None, 126, 126, 32)	896	input_shadow[0][0]
max_pooling2d (MaxPooling2D)	(None, 63, 63, 32)	0	conv2d[0][0]
max_pooling2d_2 (MaxPooling2D)	(None, 63, 63, 32)	0	conv2d_2[0][0]
conv2d_1 (Conv2D)	(None, 61, 61, 64)	18496	max_pooling2d[0][0]
conv2d_3 (Conv2D)	(None, 61, 61, 64)	18496	max_pooling2d_2[0][0]
max_pooling2d_1 (MaxPooling2D)	(None, 30, 30, 64)	0	conv2d_1[0][0]
max_pooling2d_3 (MaxPooling2D)	(None, 30, 30, 64)	0	conv2d_3[0][0]
flatten (Flatten)	(None, 57600)	0	max_pooling2d_1[0][0]
flatten_1 (Flatten)	(None, 57600)	0	max_pooling2d_3[0][0]
concatenate (Concatenate)	(None, 115200)	0	flatten[0][0] flatten_1[0][0]
output_height (Dense)	(None, 1)	115201	concatenate[0][0]

Total params: 153,985
Trainable params: 153,985
Non-trainable params: 0

Fig. 6. Model Architecture.

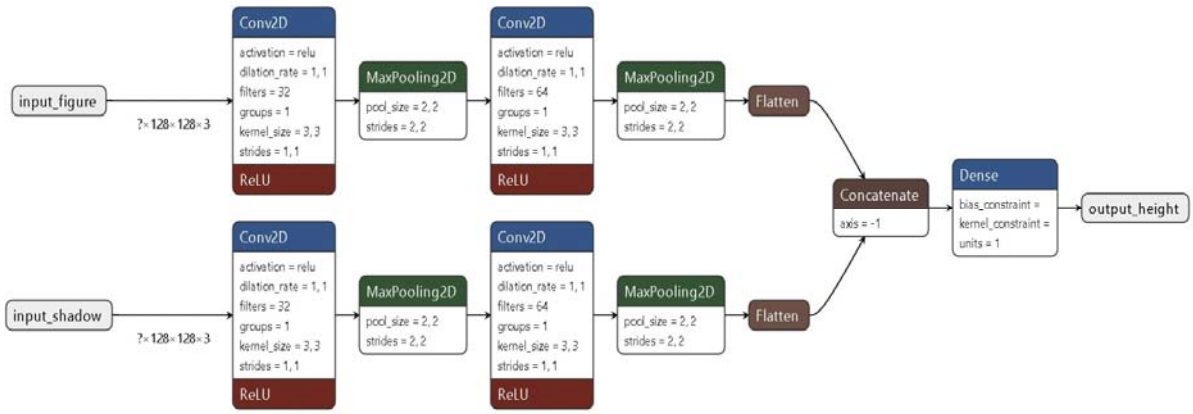

Fig. 7. General Scheme of the Proposed Model (horizontal view). (Reported using Netron. https://netron.app/).

General Scheme of the Proposed Model (horizontal view). Is a regression convolutional neural network architecture designed to process two distinct inputs: from an image and from its corresponding shadow. Each input was initially processed by two parallel convolutional layers (conv2d and conv2d_2), both conformed with filters of size 3×3 and a depth of 32, using the ReLU activation function to introduce nonlinearities.

After this first stage, the features extracted from both inputs were MaxPooling (max_pooling2d and max_pooling2d_2) with a pool size of 2×2 and strides of 2, which reduces the dimensionality obtaining the most relevant features. Next, the resulting feature

maps went through a second set of convolutional layers (conv2d_1 and conv2d_3) which extended the filter depth to 64, keeping the same kernel size and ReLU activation function. A second MaxPooling operation (max_pooling2d_1 and max_pooling2d_3) followed this layer, further consolidating the relevant features.

The data were then flattened with the flatten and flatten_1 layers in order to transform the multidimensional output into a one-dimensional vector, which is necessary for the concatenation of the data coming from the two branches of the network model. This concatenation was performed along a common axis (axis = -1), which is the last axis (dimension) in a tensor, allowing the fusion (concatenation) of the image data and their shadows into a joint representation. Finally, the network produced a single output (output_height) through a dense layer, which was configured with a single node and used a kernel of size (115200×1), suggesting that the flattened data were of significant dimension before the final operation. This output layer provided a scalar estimate, possibly related to the height of the object depicted in the image, which is a common interpretation in regression tasks with neural networks in the context of image processing.

With this knowledge, it is expected that the convolutional neural network will be able to predict the height of the shapes from their shadow. This height measurement is given in centimeters (units) using as input image only the projected shadow. Thus, it is initially expected to effectively estimate the height of the shape or object corresponding to the shadow, which is one of the challenges of the research. Moreover, there is additional information in the dataset that combined with appropriate 2D to 3D shape reconstruction algorithms it is possible to obtain the reconstruction of the shape according to the height established from the shadow.

In [21] (M. Daum, G. Dudek. 1998) the reconstruction of 3D surfaces from shadows is reported. Here, the movement of shadows that are cast is used to obtain information about the scene structure, based on the collection of images from a fixed point as an illumination source moves. As a result, a 3D scene reconstruction algorithm is obtained taking into account the trajectories of the light source.

On the other hand, in [26] (Huang, X., Gao, J., et al. 2007) their research addresses shape estimation from shading (SFS) which aim to solve a problem with few constraints to estimate the depth map from a single image, developing an example-based method to improve the accuracy of (SFS), improving the reconstruction quality from real images of different shapes by obtaining prior knowledge of their appearance for three-dimensional shape recognition.

Similarly in the proposal in [20] (E. Prados, O. Faugeras. 2006) with the shape from shading (SFS) and with a single image they estimate the 3D shape of a surface, where in diffuse surface the position of the illumination source is not known and where the reflectance map is not known.

In [15] (S. Savarese, et al. 2007), [16] (S. Savarese, et al. 2002) and [5] (R. Gouiaa, J. Meunier. 2014) they propose a new method to recover a shape from shadow (SFS). These projected shadows are relevant information about the shape of objects, discovering cavities that are not available from clues such as occluded boundaries. This method is called "occluding boundaries". According to the volume occupied by an object, it is possible to identify and sculpt the regions of the volume that present inconsistencies with the pattern observed in the shadows. In summary, their proposal is a reconstruction system to recover the shape from silhouettes and shadow carving where shadow carving is used to carve the concavities and silhouettes are used to reconstruct the initial conservative estimate of the shape of the object.

On the other hand, in [3] (D. Forsyth, A. Zisserman. 1991), he shows how mutual illumination and image irradiance within an image give rise to more complex structures.

Similarly, in [27] (Atkinson, G. A., & Hancock, E. R. 2007), research presents a novel method for the reconstruction of 3D surfaces with polarization and shading information from two views. They use Fresnel theory in image processing to obtain estimates of the surface normal. Here they emphasize how the measured pixel brightness's depend on the surface orientation combined with the refined estimates to determine the correspondence between two views of an object. In [19] (R. Liu, S. Menon, et al. 2022) they address 3D reconstruction when the object structure is partially or totally occluded. They introduce into the method projected shadows of an unobserved object, to perform the inference of the possible 3D volume to which it corresponds. The "inheritable" imaging model allows to jointly infer both the 3D shape of the object, its location and that of the illumination source.

In [19] (R. Liu, S. Menon, et al. 2022) they address 3D reconstruction when the object structure is partially or totally occluded. They introduce into the method projected shadows of an unobserved object, to perform the inference of the possible 3D volume to which it corresponds. The "inheritable" imaging model allows to jointly infer both the 3D shape of the object, its location and that of the illumination source. The advantage of this approach is the estimation of multiple 3D scenes from shadows.

The data set is organized in folders and within each subfolder, there are images corresponding to each geometric shape, the shadow, the height. It is important to point out that with the experimentation process, the length of the shadow and the angle of the light source were added to the Dataset already created, improving the accuracy of the model (see Fig. 8).

Some research such as in [10] (Sanin, C. Sanderson, B. C. Lovell. 2012) presents various methods and algorithms for detection, removal of moving shadows as of 2012, and places them in a feature-based taxonomy that is made up of four categories, chromaticity, physics, geometry, and textures. The use of tracking performance is proposed as an approach to determine shadow detection

methods and algorithms. Of the most prominent methods are the method based on chromaticity is the fastest to implement and execute, but sensitive to noise and fails when the spectral properties of objects when these are similar to those of the background and the method based on textures of large and small regions where the results are more accurate, although its computational cost is high.

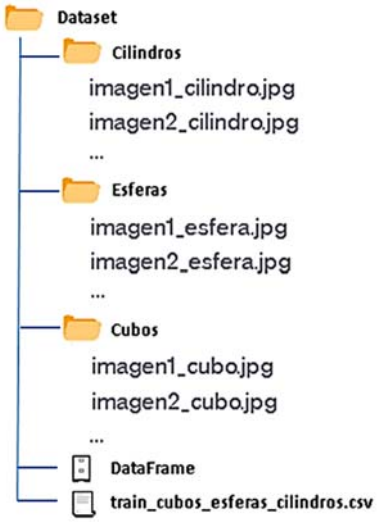

Fig. 8. Dataset Folder Structure.

Likewise, in [23] (D. Samaras, D. Metaxas. 2003) they present a method for the integration of nonlinear holonomic constraints in deformable models and its application to the problems of shape and direction estimation of the illumination source from hats. Their proposal indicates that it works when the direction of the light source is not known. They coupled the shape estimation method with a light estimation method were better shape estimation results in better light estimation and vice versa.

This research in [32] (Hoiem, D., Efros, A.A. and Hebert, M 2005) introduces a method that learns appearance-based geometric class models to estimate general geometric properties of scenes, even those that are cluttered and natural. These geometric classes define the three-dimensional orientation of regions of an image relative to the camera position. For robust estimation of the structure of a scene using a single image, the authors develop a multiple-hypothesis approach that also provides confidence measures for each geometric label.

These confidence measures have the potential to boost performance in a variety of computer vision applications. The paper not only provides a detailed quantitative evaluation of the algorithm using an exterior image data set, but also demonstrates the effectiveness of the technique in two specific practical applications: object detection and automatic scene reconstruction from a single view. The integration of geometric estimation into these processes evidences how a thorough understanding of 3D geometry can be instrumental for advances in computer vision.

Likewise, in [18] (S. H. Khan, M. Bennamoun et al. 2014) they use multiple deep convolutional neural networks (ConvNets), where their architecture consists of alternating convolution and subsampling layers. This research is focused on shadow detection, having as input a database containing in various lighting conditions with sunny, cloudy and dark environments. For this, they used a Conditional Random Field (CRF) model, which translates into predictions for each pixel of each test image, and are subsequently compared with the actual shadow masks. This enforces labeling consistency across the nodes of a gridded graph defined on the image by eliminating isolated labeling results.

In [29] (Saxena, A., Sun, M. and Ng, A.Y. 2009) they introduce an algorithm for deducing three-dimensional structures from individual still images. This algorithm outperforms previous methods, generating 3D models that are more accurate and visually superior. The approach consists of modeling the position and orientation of homogeneous regions in the image, known as superpixels, through a Markov random field (MRF). Through supervised training, the algorithm learns to estimate the parameters of the planes based on features extracted from the images and considers the interconnections between different segments of the image. Maximum a posteriori inference (MAP) of the model is efficiently executed by solving a linear program.

Unlike other models that assume scene-specific structures, such as the assumption of vertical planes with respect to the ground by Delage et al. and Hoiem et al. this algorithm is not limited to such assumptions, allowing it to generalize effectively even to scenes with significantly non-vertical structures. As a result, self-contained and aesthetically convincing 3D models have been generated for 64.9 % of the 588 Internet images tested, in contrast to the 33.1 % success rate of Hoiem et al.

2. Distribution and Structure of the Dataset

For the organization of the dataset, these were grouped according to the height measurements and their geometric shape. The dataset is made up of geometric figures such as spheres, cubes and cylinders of different heights.

The measurements were taken in centimeters, which are then handled in units, in order to facilitate the reuse of the dataset in any other measurement system (see Fig. 9).

Initially two variables (features) were defined in the dataset, the first being the type of figure or shape and the second the height of the 3D object. Subsequently, three more variables are added to the existing ones, these are the shadow, the shadow length and the angle of the light source. These variables are

within the model and were used in the learning process to improve the accuracy of the model [28] (René, J. et al. 2023).

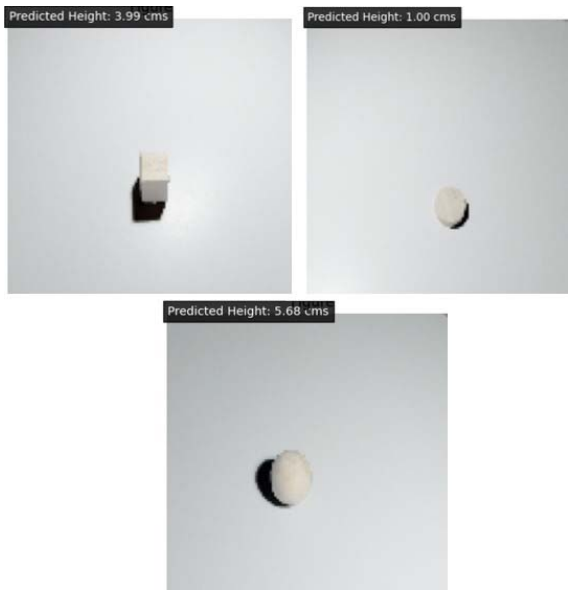

Fig. 9. Height Cylinder, Sphere and Cube.

3. Implementation and Analysis of Results

At the time of performing the network learning process, it is taken into account that it is distributed or divided into a training set (to build the model), a validation set (to test different parameters of the learning algorithm) and a test set (to evaluate the quality of the model). A suggested distribution of the data could be as follows, 80 % for the training set, 10 % for the validation set and 10 % for the test set. (Tables 2 and 3). Similarly, it could also be 70 %, 15 % and 15 % respectively.

Complementary metrics were incorporated in order to thoroughly evaluate the performance of the model throughout the training process. The model underwent training using the fit function and throughout this training period, the model's performance on the data set was continuously monitored.

Table 2. Dataset Distribution 1.

Dataset	Figures		Training	Validation	Test	Subtotal images
			80%	10%	10%	
Dataset No.1	Spheres	Spheres and Shadows	664			1.328
		Shadows	664			
	Cylinders	Cylinders and Shadows	190			380
		Shadows	190			
	Cubes	Cubes and Shadows	268			536
		Shadows	268			
Total images **(100%)**			**2.245**			**2.245**

Table 3. Dataset Distribution 2.

Dataset	Figures	Training	Validation	Test
Dataset No.2	Spheres	1.200	150	150
	Cylinders	1.200	150	150
	Cubes	1.200	150	150
Subtotal		3.600	450	450
Total images (100%)		4.500		
Dataset distribution		**80%**	**10%**	**10%**

This systematic monitoring allowed the necessary adjustments to be made to the model's hyperparameters, all with the aim of optimizing and enhancing its performance. It was trained with the dataset indicated above which allowed reaching an optimal performance on the test data and which was later validated with different data that determined its capacity to generalize new images.

The MAE value obtained is 0.24, which is interpreted to mean that on average, the model predictions have an absolute error of 0.24 units in the height prediction. In other words, on average, the model predictions deviate by 0.24 units from the actual height of the objects. Therefore, the MAE equal to 0.24 indicates the average size of the absolute errors in the predictions (see Fig. 10).

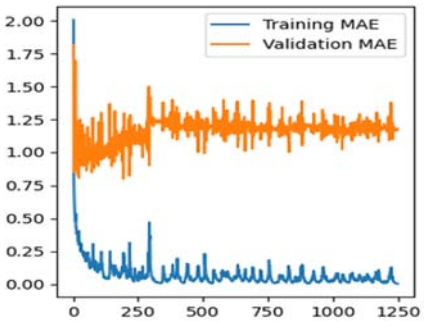

Fig. 10. Training and validation MAE for Height.

Evaluating the plot suggests that the model is learning correctly, but with room for optimization and improvement, since the model is learning as the training loss decreases over time (see Fig. 11). In the graph, the model is learning correctly, but with room for optimization and improvement, as the training loss decreases over time. It is observed that the validation loss function oscillates and, although it shows a decreasing trend, it remains relatively constant and fails to decrease at the same rate as the training loss function, which may indicate that the model has the ability to generalize to new data.

The value obtained for the MSE is 0.46 meaning that, on average, the square of the errors between the model predictions and the actual height of the shapes is 0.46. Thus, the MSE gives more weight to larger errors due to its quadratic nature. In this case, an MSE

value of 0.46 suggests that there is variability in the errors, including some errors that are larger and affect the average (see Fig. 12).

The coefficient of determination R^2 varies between 0 and 1, where 1 indicates that the model can explain all the variability in the data and 0 indicates that the model cannot explain any variability. A R^2 value of 0.929 is quite high, suggesting that the model has a good ability to explain the variability in the data.

In the tests, the Pearson correlation coefficient result is 0.8618 which is a high value and suggests that there is a strong positive linear relationship between the predictions and the actual values.

The structure of the training, validation and test dataset was structured in such a way that in addition to containing the images corresponding to each folder, it contains a flat file with the characteristics indicated above, such as the angle of the light source or the length of the shadow [28] (René, J. et al. 2023). The DataFrame structure has the following characteristics: image name (filename_image), figure name (name_figure), figure height(height_figure), shadow name (filename_image_shadow), shadow image name (name_image_shadow), shadow length (shadow_length), illumination angle (illumination_source_angle) (see Table 4).

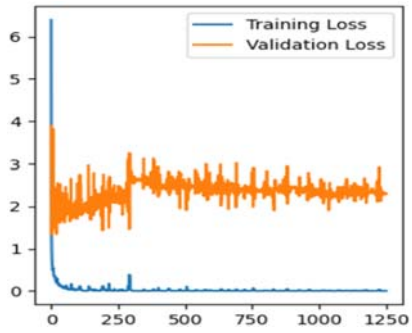

Fig. 11. Training and validation LOSS for Height.

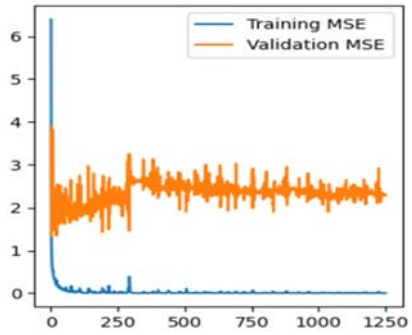

Fig. 12. Training and validation MSE for Height.

Table 4. DataFrame Structure.

filename_ image(jpg)	*name_image*	*height_figure* cilindro(cm)	*filename_image_ shadow(jpg)*	*name_image_s hadow*	*shadow_ length*	*illumination_ source_angle*
cilindro1_00001G170	cilindro	1.0	cilindro1_00001G170s	sombra_cilindro	1.0	45
cilindro1_00001G180	cilindro	1.0	cilindro1_00001G180s	sombra_cilindro	1.0	45
cilindro1_00001G335	cilindro	1.0	cilindro1_00001G335s	sombra_cilindro	1.0	45
cilindro3_00001G180	cilindro	3.0	cilindro3_00001G180s	sombra_cilindro	3.0	45
cubo7_00002G335	cubo	7.0	cubo7_00002G335s	sombra_cubo	7.0	45
cubo8_00001G80	cubo	8.0	cubo8_00001G80s	sombra_cubo	8.0	45
cubo9_00001G335	cubo	9.0	cubo9_00001G335s	sombra_cubo	9.0	45
esfera1.5_00001G100	esfera	1.5	esfera1.5_00001G100s	sombra_esfera	1.5	45
esfera1_00001G190	esfera	1.0	esfera1_00001G190s	sombra_esfera	1.0	45
esfera2.5_00002G10	esfera	2.5	esfera2.5_00002G10s	sombra_esfera	2.5	45

The graph shows the learning curve of the model during the training and validation process, plotting the loss (error) as a function of the number of epochs, as the epochs advance, the training loss curve flattens, indicating that the model is reaching a convergence point where additional adjustments to the model weights and biases have a decreasing impact on the reduction of the training loss.

On the other hand, the loss on validation stabilizes and stays within a range with no clear trend to decrease with the passage of more epochs, suggesting that the model has learned as much as it can from the training data set given the current architecture and configuration (see Fig. 13).

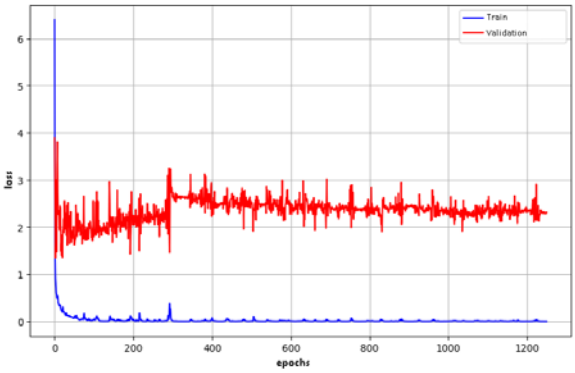

Fig. 13. Model learning curve.

4. Results and Discussion

Based on the plot the accuracy of the predictions varies over different values of actual height. For example, for real heights of around 2 to 6, the predictions are clustered around the trend line, indicating predictions with higher accuracy. Whereas, for true heights between 7 and 10, the predictions appear to be more widely dispersed, indicating lower accuracy in that range. On the other hand, predictions for actual heights of approximately 8 and 9 show considerable variability, with some points predicted well below the actual value. Despite the scatter and some significant errors, the overall trend follows the trend line, indicating that, in general, the model has some ability to predict height based on the inputs provided, albeit with a margin of error that can be substantial in some cases (See Fig. 14. Scatter plot: differences between actual vs. predicted height).

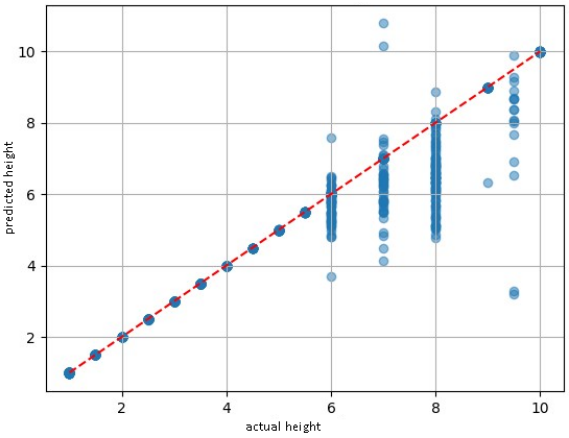

Fig. 14. Scatter plot: differences between actual vs. predicted height.

The histogram shows the distribution of the differences between the actual height and the height predicted by the proposed model. The x-axis represents the difference between the actual height and the predicted height, where negative values indicate that the predicted height was less than the actual height, positive values indicate that the predicted height was greater than the actual height, and a value of 0 would indicate an accurate prediction. On the other hand, the y-axis represents the frequency, or the number of times, that each height difference occurs. The plot shows a very high concentration of differences around 0, suggesting that most of the model predictions are very close to the actual height. This is indicative of good model performance on the height prediction task (see Fig. 15).

The box plot shows the distribution of prediction errors for three types of shapes: cylinder, cube and sphere used by the model. Analyzing the box plot shows that the model predicts the height of cylinders with low variability and good accuracy. On the other hand, for cubes, the accuracy is lower, and for spheres, although most of the predictions are accurate, there is a significant amount of large overestimation errors (the model predicts a greater height than the actual height).

These differences may indicate that the model needs additional adjustments or training to improve its accuracy in estimating height, especially for figures that are spheres. Similarly, there are outliers, which are the points outside the whiskers. For cylinders, the model is consistent, as reflected in the small and symmetric errors around the median. Similarly, the median errors for the cubes are close to zero, suggesting accuracy in the predictions, and the median errors for the spheres appear to be the furthest from zero compared to the cylinders and cubes, suggesting a tendency to systematically overestimate or underestimate (see Fig. 16).

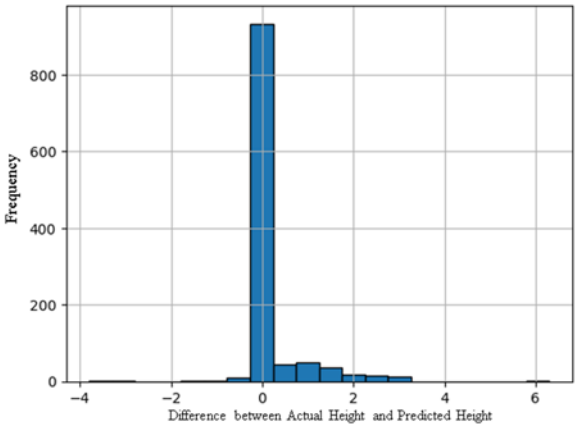

Fig. 15. Histogram of Differences between Actual Height and Predicted Height.

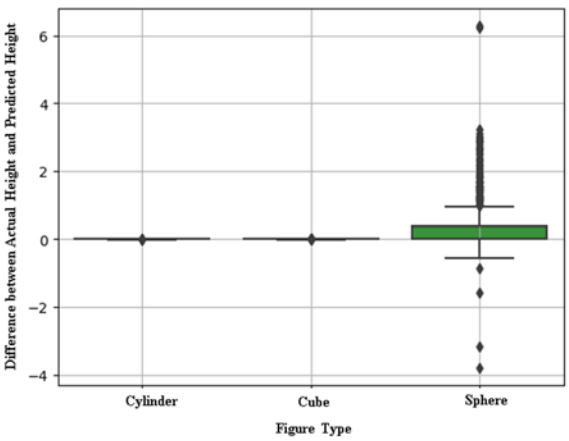

Fig. 16. Box Plot: Prediction Errors by Figure Type.

The residual plot for the model indicates a generally good fit, with consistently low errors for most samples. The plot shows data points with high residuals, which are an important source of information for improving the model; the residuals are

the difference between the actual values and the values predicted by the model for each sample (see Fig. 17).

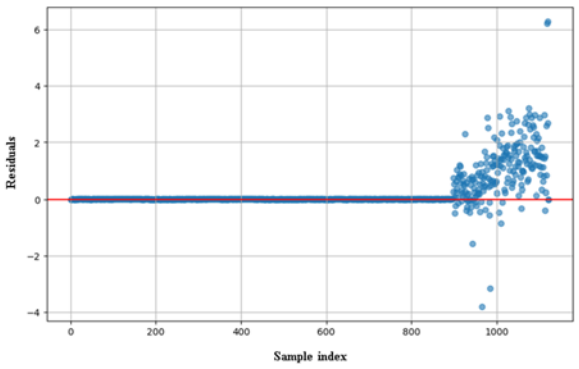

Fig. 17. Plot of residuals.

Cross-validation is a technique used to evaluate the generalizability of a model by training and evaluating it on different subsets of the original data set. In this case, a K-fold with k = 5 has been used, which means that the data has been divided into 5 subsets (folds). In each cross-validation iteration, the 4 folds are used to train the model and 1 fold is used to evaluate it. This process is repeated 5 times, each time with a different fold as the validation set. After applying cross-validation model C with 1.250 epochs, despite not having the most Epochs, outperforms the rest in terms of all metrics.

On this occasion, model C is the most outstanding in all aspects in terms of performance as well as MAE and MSE. If one were to choose a model based on these metrics alone, model C would be the choice (see Table 5).

Table 5. General results of the cross-validated model.

Model	Fold	Epochs	Loss	MAE (Mean Absolute Error)	MSE (Mean Squared Error)
Modelo A	1,2,3,4,5	600	0,625776678323745	0,516906213760376	0,625776678323745
Modelo B	1,2,3,4,5	800	0,640111273527067	0,640111273527067	0,640111273527144
Modelo C	**1,2,3,4,5**	**1.250**	**0,542674350738525**	**0,483043521642684**	**0,542674350738525**
Modelo D	1,2,3,4,5	1.600	0,571656787395477	0,486995190382003	0,571656787395477

The configuration is the same for the model, what changes is the number of epochs:

- Epochs: 1.250;
- Batch size: 32;
- Verbose: 1;
- Input shape: (128, 128, 3);
- Dataset: 2.244 images;
- k = 5: number of folds.

The model was evaluated at different epochs to determine how the number of iterations during training affects model performance. Increasing the epochs provided more opportunities for the model to learn patterns in the data, but there is also a risk of overfitting if the model is overtrained. Evolution shows that simply increasing the Epochs does not guarantee an improvement in performance. It is possible that after a certain point, more Epochs may lead to overfitting or may not bring significant improvements.

In this case, model C with 1.250 epochs has the best performance, indicating that it may be the optimal point in terms of trade-off between training time and model performance. The table shows us the cross-validation of four different models (model A, B, C and D) using 5 folds (the splits of the data for cross-validation). The results are averages of the metrics of those 5 folds (see Table 5).

The graphs represent the evolution of the performance of a machine-learning model over a number of epochs during training and validation in terms of three different metrics: the Mean Absolute Error (MAE), the Mean Squared Error (MSE) and the Loss function. The MAE and MSE metrics are regression measures that are quantifying the difference between the values predicted by the model and the actual values. The graphs show that after the first few epochs, the validation metrics (green lines) stabilize, indicating that the model is learning effectively without improving significantly after a certain point. This stabilization suggests that the model has reached its generalization ability with the validation data. In addition, the closeness between the training and validation lines on all metrics indicates that the model has consistent behavior between the training and validation data. There is no large gap that would indicate significant overfitting.

MAE values range from approximately 0.42 to 0.53, suggesting that, on average, model predictions deviate from actual values by this margin. The MSE values vary similarly to the MAE, and since the MSE penalizes larger errors more, the relatively low magnitude of these values suggests that there are no extremely large errors that are significantly affecting performance. The loss function values reflect the MSE

and are aligned with the MAE, confirming consistency in the model's performance assessment. In summary, the model performs acceptably consistently and stably throughout the validation. There are no obvious signs of overfitting and the model generalizes well to data not seen in the cross-validation. The numerical values and plots indicate that the model is reliable on different data sets, which is a desirable feature in a machine-learning model (see Figs. 18-20).

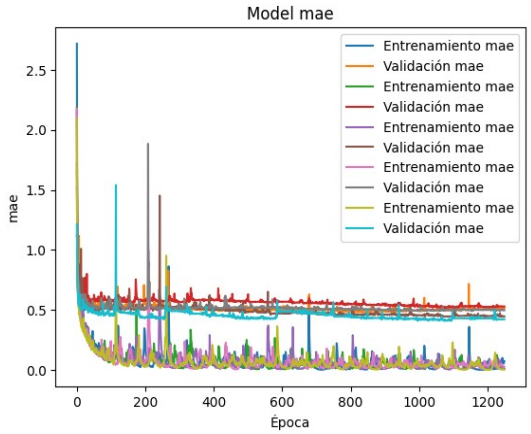

Fig. 18. Average performance in cross-validation: MAE.

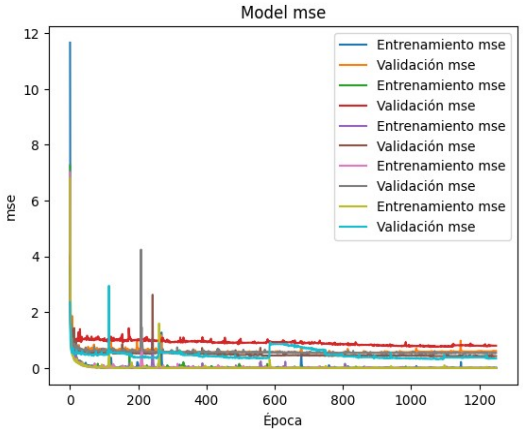

Fig. 19. Average performance in cross-validation: MSE.

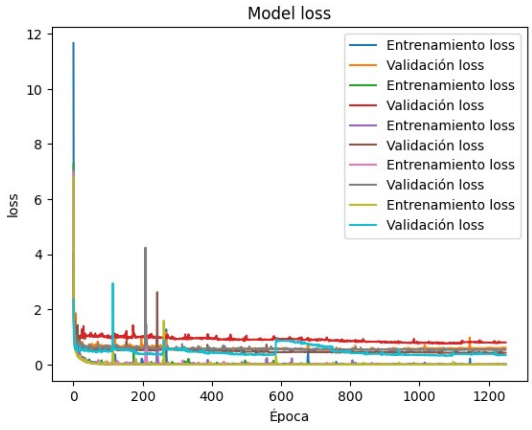

Fig. 20. Average cross-validation performance: Loss.

5. Conclusions

The areas of highest accuracy are located in the lower to middle range of true heights, where the clustering of points around the trend line suggests more reliable predictive performance. Similarly, the model can be optimized for height prediction. Optimization could include exploring more complex network architectures, incorporating regularization techniques to mitigate overfitting, making the dataset more diverse in shapes or figures and shadows.

The model seems to generalize well the relationship between the inputs and the height of the target figure. This is apparent from the histogram of differences between actual and predicted height, which shows most of the differences clustered near zero. The predictions are close to the actual values in most cases, as evidenced by the residual plot, where most of the points are close to the residual zero line.

Machine learning models are rarely complete. There is always room for improvement. Therefore, it is considered important to incorporate more data, test different neural network architectures, and even combine multiple models into one ensemble to improve performance. Also include new data for a more diverse dataset, since the quality and quantity of training data can have a large impact on model performance.

References

[1]. S. Hosseinzadeh, M. Shakeri, H. Zhang, Fast Shadow Detection from a Single Image Using a Patched Convolutional Neural Network, *Natural Science and Engineering Research Council (NSERC) through the NSERC Canadian Field Robotics Network (NCFRN) and by Alberta Innovates Technology Future (AITF)*, 2018, pp. 3124-3129.

[2]. A. Panagopoulos, S. Hadap, D. Samaras, Reconstructing shape from dictionaries of shading primitives, *Lecture Notes in Computer Science*, Vol. 7727, 2013, pp. 80-94.

[3]. D. Forsyth, A. Zisserman, Reflections on Shading, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 13, Issue 7, 1991, pp. 671-679.

[4]. T. F. Y. Vicente, C.-P. Yu, D. Samaras, Single image shadow detection using multiple cues in a supermodular MRF, in *Proceedings of the British Machine Vision Conference*, 2013, pp. 126.1-126.11.

[5]. R. Gouiaa, J. Meunier, 3D reconstruction by fusioning shadow and silhouette information, in *Proceedings of the IEEE Conference on Computer and Robot Vision (CRV'14)*, 2014, pp. 378-384.

[6]. D. J. Kriegman, et al., 3D photography using shadows in dual-space geometry, *Lecture Notes in Computer Science*, Vol. 1407, March 1999, pp. 399-414.

[7]. G. Liasis, S. Stavrou, Satellite images analysis for shadow detection and building height estimation, *ISPRS Journal of Photogrammetry and Remote Sensing*, Vol. 119, 2016, pp. 437-450.

[8]. D. J. Kriegman, P. N. Belhumeur, What shadows reveal about object structure, *Lecture Notes in Computer Science*, Vol. 1407, March 1998, pp. 399-414.

[9]. S. Mohajerani, P. Saeedi, CPNet: A context preserver convolutional neural network for detecting shadows in single RGB images, in *Proceedings of the IEEE 20th International Workshop on Multimedia Signal Processing (MMSP'18)*, 2018, pp. 1-5.

[10]. A. Sanin, C. Sanderson, B. C. Lovell, Shadow detection: A survey and comparative evaluation of recent methods, *Pattern Recognition*, Vol. 45, Issue 4, 2012, pp. 1684-1695.

[11]. D. C. Knill, P. Mamassian, D. Kersten, Geometry of shadows, *Journal of the Optical Society of America A*, Vol. 14, Issue 12, 1997, 3216.

[12]. E. Salvador, A. Cavallaro, T. Ebrahimi, Cast shadow segmentation using invariant color features, *Computer Vision and Image Understanding*, Vol. 95, Issue 2, 2004, pp. 238-259.

[13]. S. Hosseinzadeh, M. Shakeri, H. Zhang, Fast shadow detection from a single image using a patched convolutional neural network, in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'18)*, 2018, pp. 3124-3129.

[14]. D. S. Kim, M. Arsalan, K. R. Park, Convolutional Neural Network-Based Shadow Detection in Images Using Visible Light Camera Sensor, *Sensors*, Vol. 18, Issue 4, March 2018, 960.

[15]. S. Savarese, et al., 3D reconstruction by shadow carving: Theory and practical evaluation, *International Journal of Computer Vision*, Vol. 71, Issue 3, 2007, pp. 305-306.

[16]. S. Savarese, H. E. Rushmeier, Implementation of a shadow carving system for shape capture, in *Proceedings of the First International Symposium on 3D Data Processing Visualization and Transmission*, June 2002, pp. 12-23.

[17]. S. A. Shafer, T. Kanade, Using shadows in finding surface orientations, *Computer Vision, Graphics and Image Processing*, Vol. 22, Issue 1, 1983, pp. 145-176.

[18]. S. H. Khan, M. Bennamoun, F. Sohel, R. Togneri, Automatic Feature Learning for Robust Shadow Detection, in *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'14)*, September 2014, pp. 1939-1946.

[19]. R. Liu, S. Menon, C. Mao, D. Park, S. Stent, C. Vondrick, Shadows Shed Light on 3D Objects, *arXiv Preprint*, 2022, arXiv:2206.08990.

[20]. E. Prados, O. Faugeras, Shape from shading, in Handbook of Mathematical Models in Computer Vision (N. Paragios, Y. Chen, Ol. Faugeras, Eds.), *Springer*, May 2006, pp. 375-388.

[21]. M. Daum, G. Dudek, On 3-D surface reconstruction using shape from shadows, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'98)*, 1998, pp. 461-468.

[22]. A. Varol, A. Shaji, M. Salzmann, P. Fua, Monocular 3D Reconstruction of Locally Textured Surfaces, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 34, Issue 6, 2012, pp. 1118-1130.

[23]. D. Samaras, D. Metaxas, Incorporating Illumination Constraints in Deformable Models for Shape from Shading and Light Direction Estimation, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 25, Issue 2, 2003, pp. 247-264.

[24]. R. Hintze, B. Morse, Shadow Patching: Guided Image Completion for Shadow Removal, in *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV'19)*, 2019, pp. 1999-2008.

[25]. M. W. Tao, P. P. Srinivasan, S. Hadap, S. Rusinkiewicz, J. Malik, R. Ramamoorthi, Shape Estimation from Shading, Defocus, and Correspondence Using Light-Field Angular Coherence, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 39, Issue 3, 2017, pp. 546-560.

[26]. X. Huang, J. Gao, L. Wang, R. Yang, Examplar-based Shape from Shading, in *Proceedings of the 6th International Conference on 3-D Digital Imaging and Modeling (DIM'07)*, Montreal, Canada, 2007, pp. 349-356.

[27]. G. A. Atkinson, E. R. Hancock, Shape Estimation Using Polarization and Shading from Two Views, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 29, Issue 11, 2007, pp. 2001-2017.

[28]. J. René, *et al.*, Estimation of Height of a Shape a 2D Image from its Shadow using Neural Networks, in *Proceedings of the 5th International Conference on Advances in Signal Processing and Artificial Intelligence (ASPAI''23)*, 2023, pp. 1-8.

[29]. A. Saxena, M. Sun, A. Y. Ng, Make3D: Learning 3D Scene Structure from a Single Still Image, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 31, Issue 5, 2009, pp. 824-840.

[30]. X. Huang, et al., Examplar-based shape from shading, in *Proceedings of the 6th International Conference on 3-D Digital Imaging and Modeling (3DIM'07)*, 2007, pp. 349-356.

[31]. M. S. U. Khan, et al., Three-Dimensional Reconstruction from a Single RGB Image Using Deep Learning: A Review, *Journal of Imaging*. Vol 8, Issue 9, 2022, 225.

[32]. D. Hoiem, A. A. Efros, M. Hebert, Geometric context from a single image, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV'05)*, 2005, pp. 654-661.

[33]. N. Wang, Y. Zhang, Z. Li, Pixel2Mesh – Generating Meshes from Single RGB Images, in *Proceedings of the European Conference on Computer Vision (ECCV'18)*, 2018, pp. 52-67.

[34]. A. Yuniarti, N. Suciati, A review of deep learning techniques for 3D reconstruction of 2D images, in *Proceedings of the 12th International Conference on Information & Communication Technology and System (ICTS'19)*, 2019, pp. 327-331.