ARGUMENT
BIANNUAL PHILOSOPHICAL JOURNAL
e-ISSN 2084–1043 p-ISSN 2083–6635
Published online: 01.09.2023
Vol. 13 (1/2023) pp. 13–28
www.argument-journal.eu

Distance estimation using artificial neural networks: architectures, capabilities and limitations

Tomasz HACHAJ*

ABSTRACT

The ability to judge distances using vision is an extremely important skill that greatly facilitates exploration of one's immediate environment. Most commonly, spatial vision is associated with stereo vision. Although human eyes also act as a stereo vision system, we can perform a simple experiment by covering one eye and then look at our surroundings: even though we are now observing the world through a single "sensor" we can still judge which objects are closer and which are further away. Though we can also employ a slight change in viewing perspective to improve our sense of distance, this is not necessary and using even one eye and standing still we are able, through the experience we have gained, to correctly estimate the distances between the objects we can see. Also when we look at photographs although the images are two-dimensional, we are able to estimate the distance portrayed in them. In recent years many solutions based on machine learning methods and deep neural networks have been developed that can mimic this process. In particular, encoder-decoder architectures are effective in this task which allows a robot single-frame depth estimation. However, these solutions still have some limitations, which constitute a challenge for researchers and engineers. This paper will discuss the challenges faced by such architectures based on the author's experience in the practice of developing deep learning-based single-frame depth estimation algorithms.

KEYWORDS

single-frame depth estimation; encoder-decoder; deep learning; typical errors; limitations

* Associate professor at Department of Applied Computer Science, AGH University of Krakow, Poland. E-mail: thachaj@agh.edu.pl.

DOI: 10.24917/20841043.13.1.01

1. INTRODUCTION

The ability to judge distances using vision is an extremely important skill that greatly facilitates the exploration of the surrounding environment and in the case of animals even determines survival (Vallar & Maravita, 2009). Most commonly, spatial vision is associated with stereo vision. Stereo vision involves integrating the results of observations from two independent vision sensors, whose parameters and mutual position are known. Although human eyes also act as a stereo vision system, we can perform a simple experiment involving covering one eye and then look at our surroundings: even though we are now observing the world through a single sensor we can still judge which objects are closer and which are farther away. Similarly, when we look at photographs, images or a picture in three-dimensional visualizations projected onto a computer screen even though the images are two-dimensional, we can estimate the distance portrayed in them.

Human space perception is strongly correlated with the environmental geometry around the moving eye (Gordon, 1965). We know that the interpretive scaling of visual angle is a key factor in size, distance, and motion estimation. We gain knowledge over the practice of interpreting visual stimuli gradually in the process of development. It is known that perception is a complex process, where prior knowledge exerts a fundamental influence over what we see (Sciutti *et al.*, 2014). Research works indicate that even newborns readily turn toward visual stimulation, indicating that primitive localization systems operate at birth (Muir, Humphrey, & Humphrey, 1994; Neil *et al.*, 2006). It is suggested that structural maturation of the visual cortex at this age results in an increase of visual spatial integration and pattern analysis. Reviews on human functional neuroimaging that have investigated the reference frames used in different cortical regions for representing spatial locations of objects can be found in papers of Gaspare Galati *et al.* and Russell Epstein and Chris Baker (Galati *et al.*, 2010; Epstein & Baker, 2019). Visual space perception translates into motion path planning and the memorized spatial position of objects can be used by humans even when they close their eyes (Loomis *et al.*, 2002). The findings suggest that specific types of spatial perception increase with age (Ishak & Haymaker, 2018) and that adults make more consistent judgments of distance than children.

1.1. Distance estimation vs SLAM

The ability to perceive the environment and self-locate in it can be formulated as a SLAM (simultaneous localization and mapping) — type problem known from robotics. SLAM problems require multiple reference frames to

estimate the location of objects in the space surrounding the observer (Galati *et al.*, 2010). Solutions such as ORB-SLAM (Mur-Artal, Montiel, & Tardos, 2015) and its many variants (Mur-Artal & Tardos, 2017; Campos *et al.*, 2021; Mazurek & Hachaj, 2021) allow for a rapid estimation of the position of the observer as well as a mapping of its surroundings, but do not have the ability to learn (optimize) the parameters of the filters processing the input image. These methods use feature detectors (most often being the particular edges of objects) generating a sparse points cloud, which has a limited ability to accurately estimate the distance of individual objects in the image.

In order to generate a dense point cloud that would allow for the evaluation of distances to all and a not just a specific object visible in the image, a more general method is needed, taking into account not only the contours of the objects but also the texture of the surface and the proportions of the characteristic objects. Moreover, it would also be useful to eliminate the need for observer movement, so that these methods could also be used for single images rather than entire video sequences. Filters that could model such complex image processing and analysis would be difficult to design manually. Convolutional neural networks are therefore most commonly used for this purpose.

1.2. Single-frame depth estimation

In recent years many solutions based on machine learning methods and deep neural networks have been developed that allow such distance estimates to be made (Mertan, Duff, & Unal, 2022; Zhao *et al.*, 2020; Ming *et al.*, 2021; Laga *et al.*, 2022). In particular, encoder-decoder architectures (E–D) (Guo, Wang, & Wang, 2019) are effective in this task which allows a robot single-frame depth estimation. However, these solutions still have some limitations, which are a challenge for researchers and engineers.

1.3. Novelty of this paper

In this paper I will discuss the challenges faced by single-frame depth estimation architectures based on the author's experience in the practice of developing deep learning-based single-frame depth estimation algorithms. An interesting piece of information, which is not available in most contemporary papers, is what type of distance estimation errors occur most often, and whether there are common errors among popular architectures. If these errors are common in different architectures, it means that they constitute general limitations of typical encoder-decoder models. If a certain type of error occurs in a specific model, it could be either a limitation of that architecture or the effect of

specific training. I have evaluated five different state-of-the-art network architectures that were published in the last four years. They differ significantly in the number of parameters (weights) (see Table 1), ranging from 2.1 M to over 100 M. I conducted a comparison by generating a depth video for 121 recordings collected during drone flights under laboratory conditions. Each of the algorithms I have used already has its own numerical quality assessment using well-known benchmarks, however to my knowledge they have not yet taken a comprehensive look at the typical errors and limitations that exist between the different models and that are evident in the generated recordings. The resulting conclusions provide important insights into the capabilities and limitations of modern E–D based single-frame depth estimation algorithms.

2. MATERIALS AND METHODS

In this section, I will discuss the neural networks I used in the experiment and the training sets on which they were learned and evaluated.

2.1. Data set

The NYU-Depth V2 data set (Silberman *et al.*, 2012) is one of the most popular data sets used by almost every depth prediction algorithm for both training and validation. It is comprised of 1449 densely labelled pairs of aligned RGB and depth images recorded by cameras from the Microsoft Kinect (Han *et al.*, 2013). I have used NYU-Depth V2 to train and numerically validate each neural network.

A new data set consisting of 112 recordings made with the UAV's (unmanned aerial vehicle) camera was proposed in the paper (Hachaj, 2022). To acquire those recordings the drone moved through an indoor space (laboratory room) that was 7.20 m long, around 2 m wide, and around 4 m high. The room was artificially lit, and the windows were covered with blinds. The room had various furniture such as desks, boxes, chairs, etc. There were various types of obstacles in the vehicle's path, both moving and static. In this work, I use this collection to visually assess from it whether the tested algorithms correctly detect the mutual position of the objects (which one is closer and which one is farther away) as well as whether any objects are missed during distance estimation. These observations are not intended to numerically assess the quality of the measurement made by the algorithm, but to detect typical errors that are visible to the human naked eye. The results obtained in this way are important because the human observer is specialized in studying the mutual distances of objects from each other and in detecting objects that are relevant from the

perspective of movement. Data was gathered with Tello UAV that is lately a popular choice in various research (Subash *et al.*, 2020). The recording was calibrated using a pinhole camera model (Zhang, 2000).

2.2. Selected state-of-the-art networks architectures

A large number of contemporary networks used for depth image prediction use U-net like architectures (Ronneberger, Fischer, & Brox, 2015). These models are used, for example, in object segmentation (Yao *et al.*, 2020; Siddique *et al.*, 2021) and reconstruction (Tang *et al.*, 2022). By using so-called skip-connections, U-nets process data from several resolutions at individual decoder layers.

Figure 1 shows the architectures of four E-D that have been proposed over the past four years for single-frame depth estimation. These are SPEED (Papa *et al.*, 2022) (the source code is available at: https://github.com/lorenzopapa5/SPEED), proposition of Ibraheem Alhashim and Peter Wonka (Alhashim & Wonka, 2018a; https://github.com/ialhashim/DenseDepth), MIDAS (Ranftl *et al.*, 2022; https://github.com/isl-org/MiDaS) and E–D proposed by Tomasz Hachaj (Hachaj, 2022; https://github.com/browarsoftware/tello_obstacles). For the purpose of this work, I also prepared an implementation of Alhashim and Wonka algorithm (Alhashim & Wonka 2018a) in which I removed the fourth skip-connection, which resulted in a reduction of the number of network weights by nearly half. I have labelled this network in Table 1 as Alhashim and Wonka (small version). For evaluation purposes I used pretrained network weights or, if these were not available, I did the training on a NYU-Depth V2 data set using the Adam optimizer (Kingma & Ba, 2015).

The above networks differ significantly in the number of parameters (weights) (see Table 1), ranging from 2.1 M to over 100 M. The number of network weights translates into the speed of the network architecture and its hardware requirements. As is usually the case in scientific papers, each of those networks is reported to be effective and useful in practical applications. An interesting piece of information, which is not available in any of these papers except (Hachaj, 2022) is what type of distance estimation errors occur most often, and whether there are common error types in all these architectures. If these errors are common to different architectures, it means that they constitute general limitations of typical U-net based encoder-decoder models. If a certain type of error occurs in a specific model, it could be either a limitation of that architecture or the effect of a training algorithm choice.

The construction of the encoder of each network I consider in this research is based on a well-known convolutional feature extractor (DenseNet — Huang *et al.*, 2017; ResNet — He *et al.*, 2016), which serves as the backbone of the

network. An important difference is the number of skip connections and the decoder layer architectures (see bottom of Figure 1).

Another important aspect critical to the quality of the network's performance is its training procedure and the loss function. Let us assume that is a ground truth depth image and is an image with predicted depth values. Index $i \in [1, ..., n]$.

A very common loss, used for example, in the methods described by Alhashim, Wonka and Hachaj (Alhashim & Wonka, 2018a; Papa *et al.*, 2022; Hachaj, 2022) is a three-element function composed of the following components:

- Point-wise depth loss for image index i:

$$l_{depth,i} = mean(|\widehat{a_i} - a_i|) \quad (1)$$

where mean(X) is the averaged value of matrix X elements.

- Edge-wise loss for image index i:

$$l_{edges,i} = mean(\left|\frac{\partial \widehat{a_i}}{\partial x} - \frac{\partial a_i}{\partial x}\right| + \left|\frac{\partial \widehat{a_i}}{\partial y} - \frac{\partial a_i}{\partial y}\right|) \quad (2)$$

- Structural similarity (SSIM) index (Z. Wang et al. 2004) loss for image index i:

$$l_{ssim,i} = mean(\frac{1 - simm(\widehat{a_i} - a_i, maxdepth)}{2}, 0,1) \quad (3)$$

where clip (x, 0, 1) is an element-wise value clipped to the range (0, 1) and the maximum depth is the maximal value of the depth pixel in the image.

The final loss function becomes:

$$l_i = w_1 \cdot l_{ssim,i} + w_2 \cdot l_{edges,i} + w_3 \cdot l_{depth,i} \quad (4)$$

where $w_1 = 1$, $w_2 = 1$, $w_3 = 0.1$ (Alhashim & Wonka, 2018b).

Image augmentation during the training consists of colour modification and mirroring.

The above-defined loss function is, of course, not the only effective way to optimize the network. René Ranftl (Ranftl *et al.*, 2022) uses the following loss: let us define $l_{ssi,i}$ as the absolute deviations term, trimming the 20% largest residuals in every image:

$$l_{issi,i} = \rho_{\text{mae80\%}}(\widehat{a_i} - a_i) \tag{5}$$

Where $\rho_{\text{mae80\%}}$ is a mean of the 80% smallest residuals between pixels in $\widehat{a_i}$ and a_i.

Another component of this loss is the gradient matching term:

$$l_{reg,i} = \frac{1}{M}\sum_{K=1}^{K}(\left|\nabla_x R_i^k\right| + \left|\nabla_y R_i^k\right|) \tag{6}$$

Where $R_i = \widehat{a_i} - a_i$ and R^k denotes the difference of the disparity maps at scale k.

Finally, the loss used by Ranftl (Ranftl *et al.*, 2022) is of the form:

$$l_{2,i} = l_{issi,i} + \alpha \cdot l_{reg,i} \tag{7}$$

Where α = 0.5.

2.3. Evaluation metrics

There are four state-of-the-art metrics that are used for the numerical evaluation of depth estimation performance (Eigen, Puhrsch, & Fergus, 2014):

- average relative error (lower is better):

$$REL = \frac{1}{n} \cdot \sum_i^n \frac{|\widehat{a_i} - a_i|}{a_i} \tag{8}$$

- root mean squared error (lower is better):

$$RMS = \sqrt{\frac{1}{n}\sum_i^n(\widehat{a_i} - a_i)^2} \tag{9}$$

- average ($\log_{10}$) error (lower is better):

$$log_{10} = \frac{1}{n}\sum_i^n|log_{10}(\widehat{a_i}) - log_{10}(a_i)| \tag{10}$$

- threshold accuracy (higher is better):

$$\delta_j = \frac{\#(\max\left(\frac{\widehat{a_i}}{a_i}, \frac{a_i}{\widehat{a_i}}\right) < t_j)}{n} \tag{11}$$

where $t_1 = 1.25$, $t_2 = 1.25^2$, $t_3 = 1.25^3$.

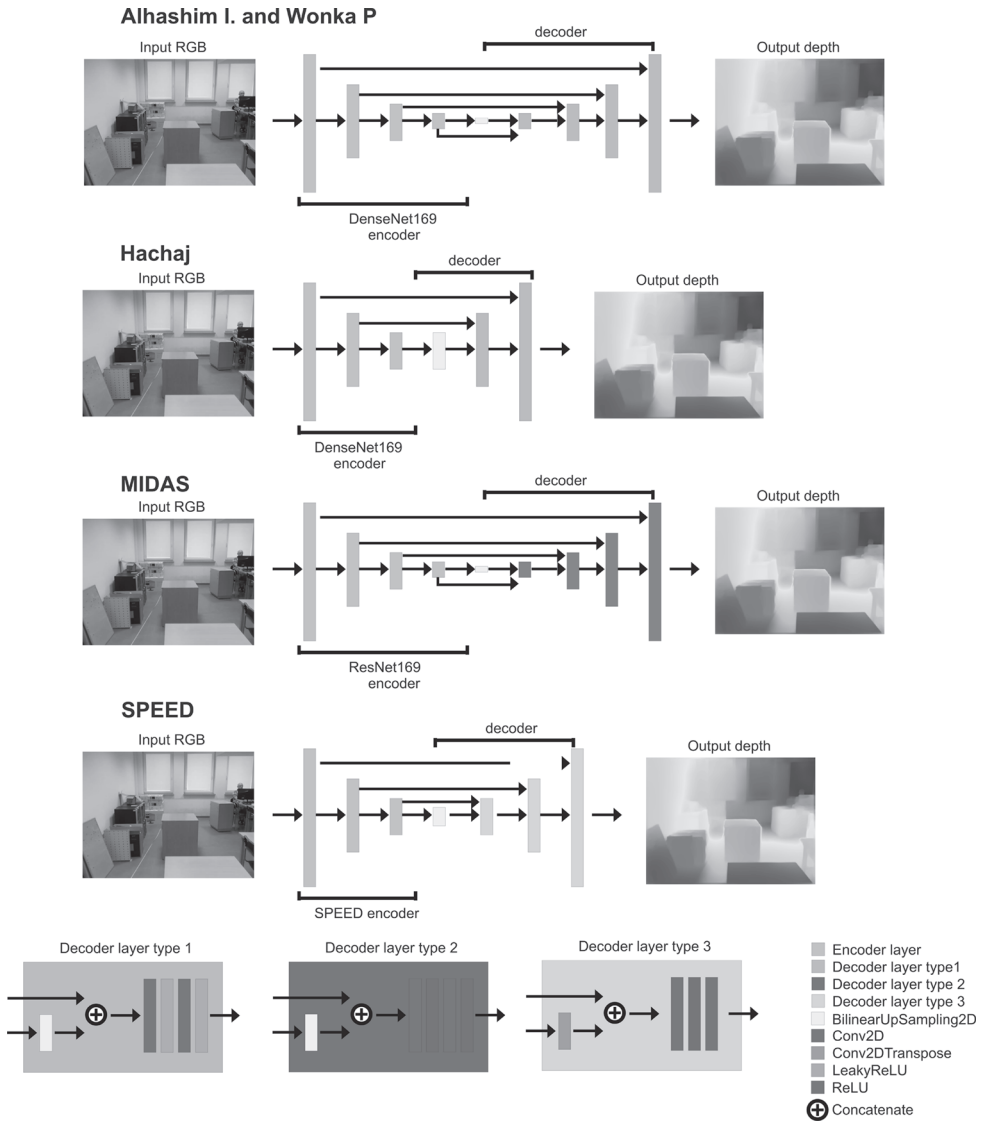


Figure 1: The architectures of the neural networks used in the study in this work: SPEED (Papa *et al.*, 2022), Alhashim and Wonka (Alhashim & Wonka, 2018b), MIDAS (Ranftl *et al.*, 2022) and Hachaj (Hachaj, 2022). Alhashim and Wonka (small version) is the same as proposed by Alhashim and Wonka (Alhashim & Wonka, 2018b), only it does not have a fourth skip-connection.

Table 1: Comparison of the performance of various depth estimation neural networks on the NYU-Depth-v2 data set. The results are reported from the original papers. The second column shows the numbers of parameters in millions (M).

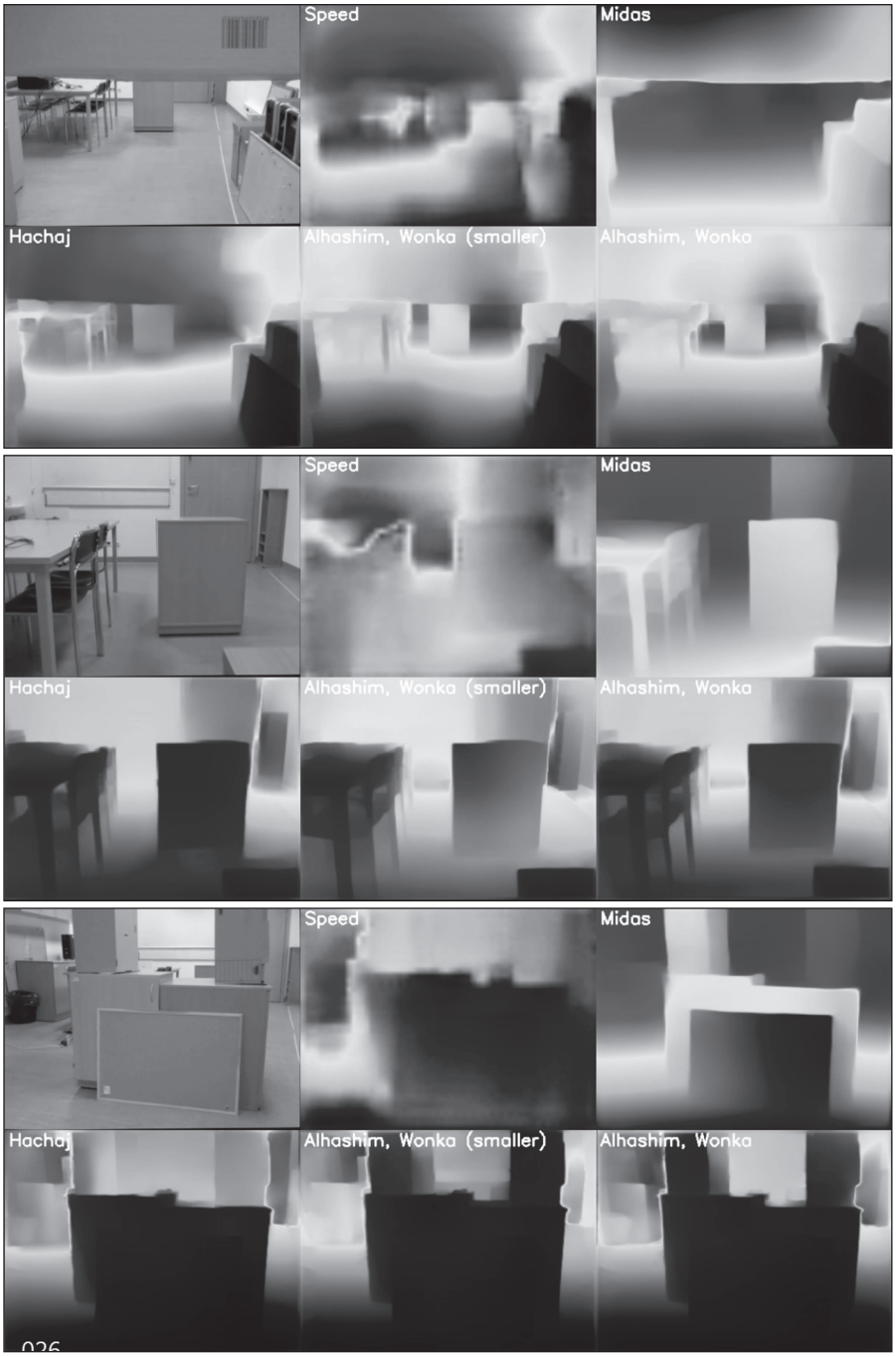
Method	#Params (M)	$\delta 1 \uparrow$	$\delta 2 \uparrow$	$\delta 3 \uparrow$	REL $\downarrow$	RMS $\downarrow$	log10 $\downarrow$
MIDAS	> 100	0.832					
SPEED	2.1	0.783			0.158	0.566	
Alhashim I. and Wonka P.	42.8	0.846	0.974	0.994	0.123	0.465	0.053
Alhashim I. and Wonka P. (small ver.)	21.5	0.841	0.972	0.993	0.130	0.567	0.054
Hachaj	6.3	0.819	0.965	0.992	0.139	0.587	0.059

3. RESULTS

The numerical performances of networks described in Section 2.2 tested on NYU-Depth V2 detest are shown in Table 1. Then I generated distance estimates for each of the 112 UAV's video recordings introduced in Section 2.1 using all five networks. Each of these video recordings was then viewed to determine if there were any distortions in the objects' mutual positions and distance misjudgements that were visible to the naked eye. The confrontation of the metrics from Table 1 with the results of the visualization of distance measurements made by the tested algorithms provides an opportunity to evaluate the suitability of the algorithm's performance in specific scenarios. I will discuss these in the next section.

Example visualizations showing the misjudgements of distances by a single or each of the networks are shown in Figure 2 and 3. Figure 4 presents a 3D reconstruction of the observed scene using the point cloud obtained from the neural network (Hachaj, 2022). Each of the tested algorithm introduced perturbations similar to those seen in Figure 4. Due to this I have in this paper presented point cloud visualizations from only one network in order not to be redundant.

Figure 2: Example depth estimations of the tested algorithms. A discussion can be found in Section 4.



(a)

(b)

(c)

Figure 3: Another example depth estimations of the tested algorithms. A discussion can be found in Section 4.

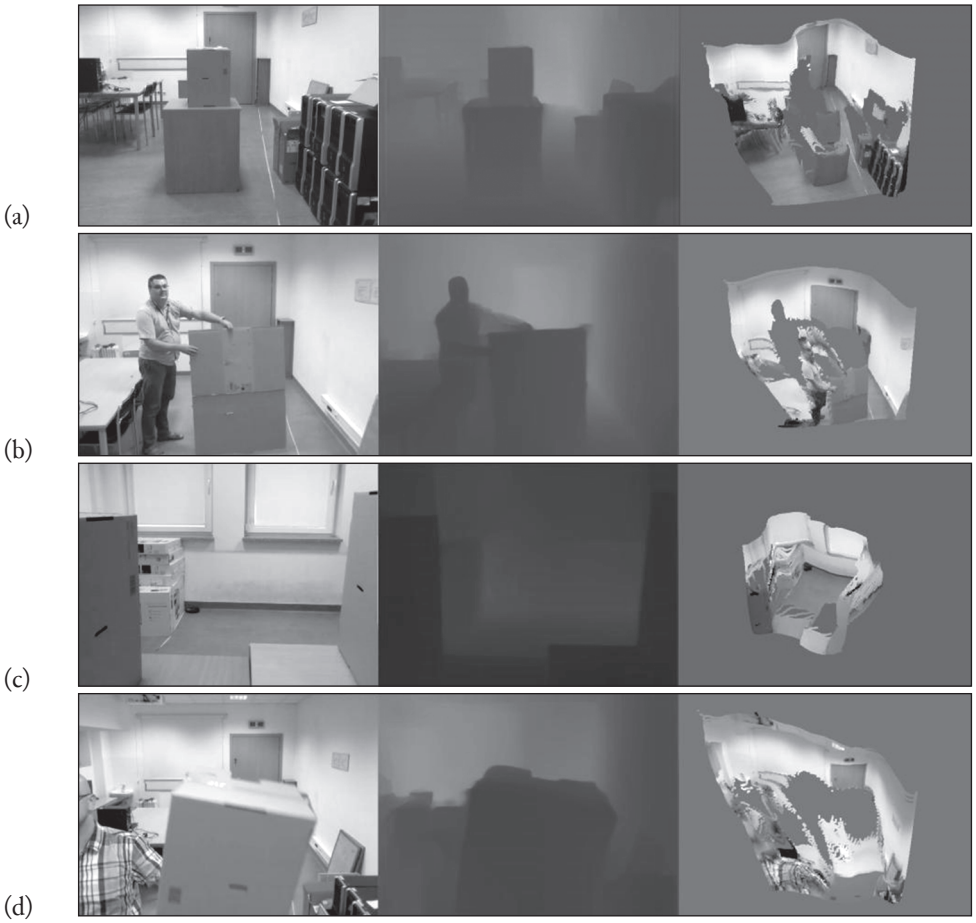


Figure 4: 3D reconstruction of the observed scene using the point cloud obtained from the neural network (Hachaj, 2022).

4. DISCUSSION

The results in Table 1 are not sufficient to evaluate the usefulness of a certain network in practice. It is also difficult to interpret the results of metrics (8)–(11). It is not entirely clear what is the direct translation of the value of the evaluation metric into the quality and usefulness of the distance estimate. The evaluation on 112 video recordings proves that Speed algorithm in practice does not work (!), while its results in Table 1 are not much lower than the evaluation values of the other methods. MIDAS seems to be visually much better than the other tested methods (it gives the smoothest results), but its σ_1 value is not the highest among all the tested methods. What is more based

on the observations made during the experiments all networks are able to judge the distances of objects located at a minimum distance of approximately 20–30 cm from the camera. If the objects are closer, networks do not work properly and recognize the objects' textures as separate objects. The larger the network is (the more weights it has); the better it is at estimating the mutual position of various objects visible in the scene. For example, only MIDAS in Figure 2 (a) coped with detecting the object that is at the top of the image. The SPEED network provides virtually no visually useful results. None of the networks is error-free. As can be seen in Figure 3 (c) no architecture could correctly locate and estimate the distance from two boxes standing parallel to each other with free space between them. In the case of Figure 2 (b), MIDAS misjudged the distance by concluding that the hand of the person holding the package was in front of the package. This error was equally not avoided by the other architectures. In some cases, the networks fail to estimate distances from flat surfaces in solid colour (see Figure 4 (a)). Also distance estimates are unstable and fluctuate depending on micro-interference on the video sensor.

Deep single-frame E–D cannot be used as an accurate method to make distance measurements. Each network returns varying values that may be non-linearly scaled from reality. The most important limitation is that an up-to-date single frame depth estimation networks cannot precisely estimates distances but merely determinate which objects are closer and which are further away.

Based on the research I have done, it can be assumed that most often the larger the network is, the more accurate the distance estimations are. In practice, this affects the calculation time. At the moment, it seems that MIDAS architecture has reached the limit of distance estimation quality for a U-net type architecture. In order to make a qualitative leap it would probably be necessary to use a different network structure than the deep encoder-decoder employed.

Currently, well-functioning networks have so many parameters that it is difficult to miniaturize them for mobile devices (microcomputers) and adapt them to work with TPU co-processors such as Edge (Cass, 2019) in order to make them operational in real time. If the number of neural network weights could be reduced, the energy expenditure required to operate distance-estimating networks could be reduced.

5. CONCLUSION

In summary, modern single-frame depth estimation algorithms based on E–D architecture are effective for estimating the mutual distance of objects, but they require relatively high computing power to operate in real time. They cannot be used to make accurate distance measurements. None of the tested architectures always works correctly. A recurring error is the misjudgement of the distance of

objects located at the corners of the image as well as the misjudgement of the distance of objects composed of several solids lying on top of each other or close to each other. The occurrence of this type of error is a common feature of all tested algorithms. Reduction of the number of network parameters results in a reduction of accuracy right up to the complete cessation of functionality.

An important challenge for the future is to propose new neural architectures and new training methods and loss functions that would allow for the miniaturization of the network. It also seems necessary to propose new network architectures that would eliminate errors that occur in each of the tested algorithms. An assumption can be made that merely increasing the number of weights of E–D convolutional networks may not be sufficient.

BIBLIOGRAPHY

Alhashim, I. & Wonka, P. (2018a). High quality monocular depth estimation via transfer learning". *arXiv e-prints*: abs/1812.11941, *arXiv:* 1812.11941, *eprint*: 1812.11941. Retrieved from: https://arxiv.org/ abs/1812.11941 (11.07.2023).

Alhashim, I. & Wonka, P. (2018b). High quality monocular depth estimation via transfer learning. *CoRR*: abs/1812.11941, *arXiv*: 1812.11941. Retrieved from: http://arxiv.org/ abs/1812.11941 (11.07.2023).

Campos, C., Elvira, R., Rodríguez, J.J.G., Montiel, J.M.M., & Tardós, J.D. (2021). ORB-SLAM3: An accurate open-source library for visual, visual-inertial and multi-map SLAM. *IEEE Transactions on Robotics*, *37*(6), 1874–1890.

Cass, S. (2019). Taking AI to the edge: Google's TPU now comes in a maker-friendly package. *IEEE Spectrum*, *56*(5), 16–17. DOI: 10.1109/MSPEC.2019.8701189.

Eigen, D., Puhrsch, Ch., & Fergus, R. (2014). Depth map prediction from a single image using a multi-scale deep network (vol. 27, pp. 2366–2374). In: Z. Ghahramani, M. Welling, C. Cortes, N.D. Lawrence, & K.Q. Weinberger (Ed.). *Advances in neural information processing systems*. Red Hook: Curran Associates, Inc. Retrieved from: https://proceedings.neurips.cc/paper/2014/file/7bccfde7714a1ebadf06c5f4cea752c1-Paper.pdf (11.07.2023).

Epstein, R. & Baker, Ch. (2019). Scene perception in the human brain. *Annual Review of Vision Science*, *5*. DOI: 10.1146/annurev-vision-091718-014809.

Galati, G., Pelle, G., Berthoz, A., & Committeri, G. (2010). Multiple reference frames used by the human brain for spatial perception and memory. *Experimental brain research / Experimentelle Hirnforschung / Expérimentation cérébrale*, *206*, 109–220. DOI: 10.1007/s00221-010-2168-8.

Gordon, D. (1965). Static and dynamic visual fields in human space perception. *Journal of the Optical Society of America*, *55*, 1296–1303. DOI: 10.1364/JOSA.55.001296.

Guo, W., Wang, J., & Wang, S. (2019). Deep multimodal representation learning: A survey. *IEEE Access*, *7*, 63373–63394. DOI: 10.1109/ACCESS.2019.2916887.

Hachaj, T. (2022). Potential obstacle detection using RGB to depth image encoder–decoder network: Application to unmanned aerial vehicles. *Sensors*, *22*(17), 1–15. DOI: 10.3390/s22176703. Retrieved from: https://www.mdpi.com/1424-8220/22/17/6703 (11.07.2023).

Han, J., Shao, L., Xu, D., & Shotton, J. (2013). Enhanced computer vision with Microsoft kinect sensor: A review. *IEEE Transactions on Cybernetics*, *43*(5), 1318–1334. DOI: 10.1109/TCYB.2013.2265378.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778. DOI: 10.1109/CVPR.2016.90.

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. (2017). Densely connected convolutional networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, *1*, 2261–2269. DOI: 10.1109/CVPR.2017.243.

Ishak, S. & Haymaker, J. (2018). Examining functional spatial perception in 10-year-olds and adults. *Perceptual and Motor Skills*, *125*, 003151251879061. DOI: 10.1177/0031512518790615.

Kingma, D.P. & Ba, J. (2015). Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Laga, H., Jospin, L.V., Boussaid, F., & Bennamoun, M. (2022). A survey on deep learning techniques for stereo-based depth estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(4), 1738–1764. DOI: 10.1109/TPAMI.2020.3032602.

Loomis, J., Da Silva, J.A., Fujita, N., & Fukusima, S.S. (2002). Visual space perception and visually directed action. *Journal of Experimental Psychology: Human Perception and Performance*, *18*, 906–921. DOI: 10.1037/0096-1523.18.4.906.

Mazurek, P. & Hachaj, T. (2021). SLAM-OR: Simultaneous localization, mapping and object recognition using video sensors data in open environments from the sparse points cloud. *Sensors*, *21*(14), 1–19. DOI: 10.3390/s21144734. Retrieved from: https://www.mdpi.com/1424-8220/21/14/4734 (11.07.2023).

Mertan, A., Duff, D.J., & Unal, G. (2022). Single image depth estimation: An overview. *Digital Signal Processing*, *123*, 103441. DOI: https://doi.org/10.1016/j.dsp.2022.103441. Retrieved from: https://www.sciencedirect.com/science/article/pii/S1051200422000586 (11.07.2023).

Minga Y., Menga, X., Fana, C., & Yu, H. (2021). Deep learning for monocular depth estimation: A review. *Neurocomputing*, *438*, 14–33. DOI: https://doi.org/10.1016/j.neucom.2020.12.089. Retrieved from: https://www.sciencedirect.com/science/article/pii/S0925231220320014 (11.07.2023).

Muir, D., Humphrey, D., & Humphrey, G. (1994). Pattern and space perception in young infants. *Spatial Vision*, *8*, 141–165. DOI: 10.1163/156856894X00288.

Mur-Artal, R., Montiel, J.M.M., & Tardos, J.D. (2015). ORB-SLAM: A versatile and accurate monocular SLAM system. *IEEE Transactions on Robotics*, *31*(5), 1147–1163. DOI: 10.1109/TRO.2015.2463671.

Mur-Artal, R. & Tardos, J.D. (2017). ORB-SLAM2: An open-source SLAM system for monocular, stereo and RGB-D cameras. *IEEE Transactions on Robotics*, *33*(5), 1255–1262. DOI: 10.1109/TRO.2017.2705103.

Neil, P.A., Chee-Ruiter, C., Scheier, C., Lewkowicz, D.J., & Shimojo, S. (2006). Development of multisensory spatial integration & perception in humans. *Developmental Science*, *9*, 454–464. DOI: 10.1111/j.1467-7687.2006.00512.x.

Papa, L., Alati, E., Russo, P., & Amerini, I. (2022). SPEED: Separable Pyramidal Pooling Encoder-Decoder for real-time monocular depth estimation on low-resource settings. *IEEE Access*, *10*, 44881–44890. DOI: 10.1109/ACCESS.2022.3170425.

Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., & Koltun, V. (2022). Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(3), 1–14.

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation (pp. 234–241). In: N. Navab, J. Hornegger, W. Wells, & A. Frangi (Ed.). *Medical image computing and computer-assisted intervention — MICCAI 2015*. Cham: Springer International Publishing.

Sciutti, A., Burr, D., Saracco, A., Sandini, G., & Gori M. (2014). Development of context dependency in human space perception. *Experimental Brain Research*, *232*, 3965–3976. DOI: 10.1007/s00221-014-4021-y.

Siddique, N., Paheding, S., Elkin, C.P., & Devabhaktuni V. (2021). U-Net and its variants for medical image segmentation: A review of theory and applications. *IEEE Access*, *9*, 82031–82057. DOI: 10.1109/ACCESS.2021.3086020.

Silberman, N., Hoiem, D., Kohli, P., & Fergus, R. (2012). Indoor segmentation and support inference from RGBD images. In: A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, & C. Schmid (Ed.). *Computer vision — ECCV 2012*. (= *Lecture Notes in Computer Science*, *7576*). Berlin–Heidelberg: Springer. DOI: https://doi.org/10.1007/978-3-642-33715-4_54.

Subash, K.V.V., Srinu, M.V., Siddhartha, M.R.V., Harsha, N.C.S., & Akkala, P. (2020). Object detection using ryze tello drone with help of mask-RCNN (pp. 484–490). In: *2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA)*. IEEE Xplore Part Number: CFP20K58-ART. DOI: 10.1109/ICIMIA48430.2020.9074881.

Tang, S., Ding, Y., Zhou, H.-W., Zhou, H. (2022). Reconstruction of sparsely sampled seismic data via residual U-net. *IEEE Geoscience and Remote Sensing Letters*, *19*, 1–5. DOI: 10.1109/LGRS.2020.3035835.

Vallar, G. & Maravita, A. (2009). Personal and extrapersonal spatial perception (vol. 1, pp. 322–336). In: G.G. Berntson & J.T. Cacioppo (Ed.). *Handbook of neuroscience for the behavioral sciences*. John Wiley & Sons Inc. DOI: 10.1002/9780470478509.neubb001016.

Wang, Z., Bovik, A.C., Sheikh, H.R., & Simoncelli, E.P. (2004). Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, *13*(4), 600–612. DOI: 10.1109/TIP.2003.819861.

Yao, Z., He, K., Zhou, H., Zhang, Z., Zhu, G., Xing, C., Zhang, J., Shao, B., Tao, Y., Sun, X., Hou, Y., Duan, M., Liu, S., Huang, L., & Zhou, F. (2020). Eye3DVas: three-dimensional reconstruction of retinal vascular structures by integrating fundus image features. *Frontiers in Optics / Laser Science*, JTu1B.22. Retrieved from: http://opg.optica.org/abstract.cfm?URI=LS-2020-JTu1B.22 (11.07.2023).

Zhang, Z. (2000). A flexible new technique for camera calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(11), 1330–1334. DOI: 10.1109/34.888718.

Zhao, C., Sun, Q., Zhang, C., Tang, Y., & Qian, F. (2020). Monocular depth estimation based on deep learning: An overview. *Science China Technological Sciences*, *63*, 1612–1627. DOI: https://doi.org/10.1007/s11431-020-1582-8.