Journal of Theoretical and Applied Information Technology
15th February 2024. Vol.102. No 3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

DEPTH ESTIMATION METHOD BASED ON RESIDUAL NETWORKS AND SE-NET MODEL

MOHAMMAD ARABIAT[1], SUHAILA ABUOWAIDA[1], ADAI AL-MOMANI[1], NAWAF ALSHDAIFAT[2], HUAH YONG CHAN [3]

[1]Department of Computer Science Faculty of Information Technology , Zarqa University, Zarqa 13100, Jordan Email:sabuoweuda@zu.edu.jo
[2] Department of Computer Science Prince Hussein Bin Abdullah, Faculty of Information Technology, Al al-Bayt University, Mafraq, 25113, Jordan.
[3]School of Computer Sciences, Universiti Sains Malaysia, 11800, Pulau Pinang, Malaysia

ABSTRACT

The objective of this study is to examine the problem of monocular depth estimation, which is essential for understanding a particular scene. The application of deep neural networks in generative models has resulted in significant progress in the precision and efficiency of depth estimations derived from a solitary image. However, most previous approaches have shown shortcomings in accurately calculating the depth barrier, resulting in less than optimal outcomes. Image restoration refers to the procedure of enhancing the quality of an image that has been degraded or has reduced clarity. This study introduces a novel and direct method that utilises the attention channel of the depth-to-depth network. This network has encoded elements that are useful for guiding the process of creating depth. The attention channel network consists exclusively of convolution layers and the Squeeze-And-Excitation Network (SENet). The deconvolution technique has the ability to produce images of excellent quality and can be efficiently taught with single-depth data. To enhance the acquisition of information and abilities in analysing the relationship between a colour value and its associated depth value in an image, we propose using a training method. The approach being suggested entails the utilisation of a color-to-depth network. This is enhanced by the inclusion of a loss function that is explicitly defined within the system. The combination with features derived from the hidden space. Regarding our attention channel network. One notable advantage of the suggested methodology is valuable due to its capacity to enhance local capabilities. To a large extent. Precise and thorough information can be Attained even in locations with intricate surroundings. The outcomes The data used in the study was derived from the NYU Depth v2 benchmark. The dataset showcases the effectiveness and durability of the proposed solution. Methodology in contrast to current cutting-edge approaches.
Keywords: *Deep Learning, Depth Estimation, Resnet-101, Features Map, NYU Depth v2.*

1. . INTRODUCTION

Lately, there has been an increasing trend towards an approach focused on inferring depth information from a single monocular image. Depth information offers vital insights, including the determination of the point where it disappears and the horizontal border, among other things, which allows for a more efficient understanding of a particular situation [1], [2], [3], [4].
Therefore, numerous applications in the field of computer vision are currently acknowledging the necessity of adding depth estimates as a fundamental prerequisite for more advanced tasks, such as 3D scene modelling and reconstruction. Depth estimation is becoming a vital component of autonomous driving systems since it can analyse the geometric structure found in acquired images. Although there have been notable improvements in predicting depth information through stereo photos and video sequences [5], [6], [7], the job of monocular depth estimation remains challenging due to the inherent uncertainty generated by its ill-posed character. In order to overcome this limitation, the first focus was on actively exploring statistical feature-based approaches. The methods in this area [8], [9] mostly focus on image segmentation as the first step, frequently utilising graph optimisation and pixel clustering algorithms. The combined fundamental distributions of

collected features from each segmented site, such as edge orientations, textures, frequency coefficients, etc., greatly aid in sensing distance. The summed distributions are used to acquire knowledge about the appropriate depth values.

On the opposite hand, various approaches have been used to accurately determine the depth values of a certain image by comparing its structural characteristics with comparable images [10], [11]. Although statistical feature-based techniques have shown potential in extracting depth information from a single monocular image, they have limitations in accurately considering the diverse variations of geometric characteristics, especially in complicated road scenarios. Lately, researchers have been inspired by the impressive accomplishments of generative models employing deep neural networks [12], [13], [14], [15]. Consequently, they have begun to investigate the utilisation of these models in estimating depth values from a single monocular image. The problem of depth estimation can be redefined as the objective of producing a depth image based on a colour image. The deep-layered architecture quickly acquires many patterns that accurately represent the depth structures of a given image, reducing the requirement for manually creating customised characteristics. To achieve this, the application of Convolutional neural networks (CNNs) with different receptive fields have become widely popular. In order to simplify the investigation of the complex relationship between colour and depth images, deep neural networks are developed utilising extensive datasets. The datasets are generated by utilising the NYU Depth v2 benchmark dataset, guaranteeing ample training data for the networks. The effectiveness of "learning" approaches allows for the accurate reconstruction of the depth map from the input colour image, even when there are complicated backdrops present.

Nevertheless, these techniques still face difficulties in the form of blurring artefacts that arise at the boundaries of the depth. This study introduces a direct and inventive method for measuring depth from a single image. The main idea of the proposed method is to efficiently guide the learning process of the color-to-depth connection by using features obtained from the latent space of the depth-to-depth network.

It is crucial to emphasise that these encoded elements accurately and succinctly represent the geometric structure, which is directly linked to the spatial layout of the image.

As a result, the accompanying gradients significantly improve the clarity of depth boundaries. The suggested approach effectively addresses depth ambiguity and reduces blurring artefacts at depth boundaries by considering the inherent characteristics of depth generation and the correlation between colour and depth values. This is demonstrated in Figure 1. Efficiently clarifying the fundamental frameworks, particularly when intricate contextual factors are present, is greatly sought after. Hence, the aim of this study is to utilise a novel depth estimation framework that can accurately determine depth from a single RGB image. The objective of this strategy is to improve the accuracy of depth estimate while also decreasing the number of factors. Therefore, this approach has the capacity to reduce computational time and demonstrate effectiveness, even when implemented on a substantial dataset. The subsequent portions of this work are organised in the following manner. Section 2 offers an extensive examination of the pertinent literature. The method for estimating depth is thoroughly explained in Section 3. The empirical results are showcased utilising a standardised dataset in Section 4. The findings are outlined in Section 5.

2 LITERATURE REVIEW

The earliest investigations concentrated on creating models that depended on statistical data obtained from a given image. [16] proposed utilising frequency features to estimate depth, both on a global and local scale.

The authors developed a probabilistic model that depends on the statistical properties of spectral magnitudes. In addition, they provided an estimated distance measurement for a certain situation. Chun et al. [17] utilised a nonlinear diffusion method to recover the ground area of indoor scenes by analysing image data statistically. Subsequently, they calculated the distance between each segment and the approximate highest point of the ground area in order to create the depth map.

Recently, there has been a strong focus on performing research to determine the ideal depth map that can be best combined with a specific colour image. This is accomplished by employing the notion of structural similarity and acquiring knowledge from previous scenarios. Karsch et al. [10] utilised the spectral characteristics of a given colour image to determine the suitable depth map. Subsequently, they optimised the calculated depth map by employing a transfer approach, such as SIFT flow. Konrad et al. [11] attempted to dynamically combine three transformation results -

color-depth, location-depth, and motion-depth - by considering their structural similarity.
Choi et al. [18] proposed the use of depth gradients as reconstruction signals in their study. These signals are then integrated into the Poisson reconstruction framework. This methodology diverges from the traditional practice of manually selecting depth values from training data. While these strategies do improve the visual quality of the estimated depth map to a significant degree, their efficacy could be enhanced when used on unknown target domains. Put simply, the precision of depth estimate is greatly dependent on the training samples utilised.

Deep learning-based methodologies. The application of generative models for depth estimation has attracted considerable attention since the progress made in categorization using deep neural networks. Eigen and colleagues. In their study, the researchers in [19] proposed a new method to build a direct correlation between colour input images and their corresponding depth maps. This was accomplished by systematically utilising deep neural networks, initially employing a rudimentary representation and subsequently enhancing it through steady refinement. The creation of the initial depth map requires the utilisation of many convolutional layers. The depth map generated is subsequently merged with the original input and fed into a secondary convolutional network to recover intricate details. Although the final output may have some blurriness due to the pooling processes used between convolution layers, the results show that the generative model has great potential in properly calculating depth maps from individual monocular images. Building upon the findings presented in the publication referenced as [19], the techniques were developed using modified algorithms. The authors [20] proposed a method to estimate the depth image by utilising segmented patches for each pixel. In order to reduce the impact of boundaries between segmented pixels, the approach employed a conditional random field that considered the variation in colours. The techniques described in references [21] and [22] utilised ResNet for the purpose of depth estimation using a single image. While these algorithms have improved the accuracy of depth estimate, they require additional information that is not limited to the depth field. Additionally, deep learning was utilised in conjunction with ground truth to determine depth estimation. In the previous study [19], the researchers considered the need of capturing both local and global contextual data to produce a precise coarse depth map. The refinement module conducts up-sampling on the coarse. The depth map is improved by iteratively adding residual learning alongside feature extraction.
The data was acquired from a prior study that utilised scale and vertical pooling techniques. Networks have demonstrated.

They exhibit remarkable proficiency in various natural language processing tasks, however they still have significant constraints. A significant constraint is the requirement for greater interpretability in the produced results. Consequently, comprehending the rationale behind the model's output or the process it employs to make decisions is challenging. Another constraint is the possibility of producing biased or unsuitable information, as the model acquires knowledge from the emergence of networks, which has enabled significant progress in the field of estimate. Extracting depth information from a single monocular image is a substantial undertaking. Previous approaches have faced difficulties in accurately explaining. The existence of a depth barrier hinders the achievement of a restoration outcome that is clear and well-defined. The article introduces a novel and direct method for detecting depth. The offered methodology entails the estimate of several parameters by utilising a solitary image, thus effectively solving the given task. The hazing artefact arises at the interface between distinct depths. The user's text should be more precise and include further details regarding the exact technical features they reference to. Kindly furnish additional details. The following section will provide an explanation.

3. PROPOSED METHOD

In this part, we outline the structure of our focus channel model and subsequently analyse the influence of the loss function in relation to our attention channel model design. The subject of network design holds great significance in the realm of computer science and telecommunications. Network architecture pertains to the arrangement and organisation of a computer network, spanning diverse components. The method that we offer has the ability to generate a depth map that is appropriate to an RGB image input, without any noticeable transitions. Figure 1 illustrates that the network consists mostly of an encoder and a

decoder. The encoder and decoder layers are connected via skip connections utilising SE-Net. The encoder employed in our analysis is RESNet-169 [23], excluding the last classification layer. This setting enables the extraction of fine details with high resolution while also downsampling the input image. The encoder employed in our investigation has been subjected to pre-training on the ImageNet dataset. The decoder in the model suggested utilises a direct up-scaling technique. This approach entails increasing the resolution of the final product from the previous layer to align with the dimensions of the output from the corresponding encoder layer after using SE-Net. The output that has been upsampled is then combined with the latter and undergoes a convolution process. The attention channel module. As mentioned before, the existence of grids distortions and unclear boundaries limits the effectiveness of depth estimation. To resolve this problem, we suggest incorporating a focus module into our network structure. This module facilitates our network in selectively distributing attention to particular pixels, enabling it to prioritise things of interest while reducing attention towards the background as shown in Figure 3.1. As a result, this method makes it easier to decrease the number of grids and improves the accuracy of edge detection. Considering the practical application circumstances, it is essential for our attention module to have a lightweight design in order to minimise the additional computational resources required, especially when they are already strained.

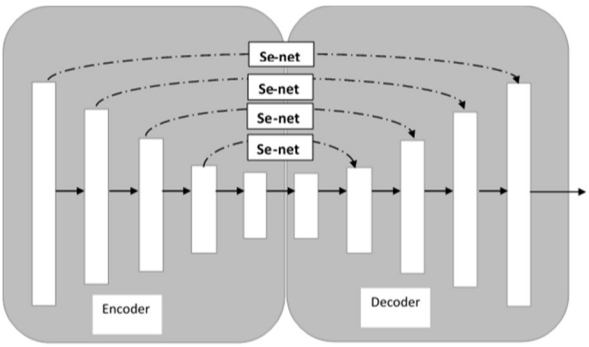

Figure 3.1. The Proposed Method Architecture

Therefore, sketching Building upon the popular inexpensive quality of SE-Net [21], we suggest incorporating a more efficient convolution attention mechanism block (SE-Net) into our model. The attention module used in our study is the SE-Net. The SE-Net selectively employs different techniques to identify and remove less important characteristics according to its weights of the feature map that is used for each layer. The SE-Net consists of five processes, as depicted in Figure 3.2.

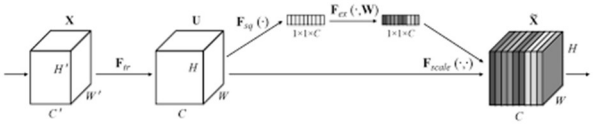

Figure 3.2. The SE-Net Architecture [21]

To incorporate content awareness techniques for assigning weights to each channel. The output of the encoder component of the ResNet is passed through the SE-Net model for each block. The objective of this phase is to acquire a linear scaler for every block. The formula of the SEN-et model is represented by the following equation. In order to comprehend the concept of the SEN-et mechanism, the SE-Net weight channel is applied to each layer of the ResNet network. This allows for a linear measurement of the block in the SE-Net network. The transfer of the ResNet network's output to the SE-Net network can be expressed as follows:

$$F_{\text{feature map}} = \sum_{i=1}^{W} \sum_{j=1}^{H} u_k(i,j)$$

The u_k item is located within the feature according to the dimensions of H × W, where H represents height and W represents width.

4 . EXPERIMENTS, RESULTS AND DISCUSSION

The effectiveness of our methodology is evaluated individually on the NYU-v2 dataset and is then compared to several baseline methods that are considered representative. Next, we present the ablation study that assesses the individual impact of each contribution. A. NYU-v2 dataset The primary emphasis of the NYU-v2 dataset is on interior environments, comprising around

120,000 frames of RGBD picture pairs. These pairings are obtained by employing both an RGB camera and the Microsoft Kinect depth sensor, enabling the simultaneous acquisition of RGB and depth data. The dimensions of the original image are 640 pixels by 480 pixels. The dataset employs an inpainting method in order to fill in missing depth values. The study adheres to the prescribed training/testing split, wherein 249 scenes are allocated for training purposes, and 215 scenes (equating to 654 photos) are designated for testing. The model is trained on a subset of 50,000 image-depth pairs, selected from a total of 120,000 pairs, as mentioned in reference [24]. A numerical evaluation of the work being proposed, the recommended model and a number of state-of-the-art models is shown in Tables I for calculating monocular depth using the NYU-V2 dataset. The comparison successfully illustrates our model's effectiveness. Once our method is used on the NYU-V2 datasets, significant improvements are seen for every parameter. We found that our method produced the best overall ranking in the NYU-V2 dataset.

Table 1. Performance Of State-Of-The-Art Architecture On Nyu Depth V2 Dataset

Models	↓ Rel	↓ Rms	↓ Log_{10}	↑ 0<1.25	↑ $0<125^2$
Eigen et al.	0.16	0.64	-	0.769	0.950
Laina et al.	0.13	0.57	0.055	0.811	0.953
Alhashim et al.	0.12	0.47	0.053	0.846	0.974
Ours	0.11	0.45	0.050	0.856	0.978

5 . CONCLUSION

The article explores the problem of monocular depth estimate from one image, considered to be a particularly difficult yet interesting case of in-depth estimation in computer vision. We present a novel strategy using an attention-based encoder-decoder network in this work. Moreover, the results of this investigation using the NYU Depth v2 data. In order to connect the encoder and decoder, the corporate lightweight attention module SE-NET is inserted into the bypass connections. The goal of integrating these elements is to pinpoint focus areas that, with the least amount of computing overhead, can minimise these visual flaws. We executed extensive trials on a dataset that serves as a benchmark and the results show that the approach we use performs better than several common standard models.

6. ACKNOWLEDGMENT

This research is funded by the Deanship of Research and Graduate Studies in Zarqa University /Jordan. The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved their research.

REFERENCES:

[1]"Borosilicate glasses containing cds/zns qds: A heterostructured com- posite with enhanced degradation of ic dye under visible-light," Chemo- sphere, vol. 286, p. 131672, 2022.

[2] C. Liu, S. Kumar, S. Gu, R. Timofte, and L. Van Gool, "Single image depth prediction made better: A multivariate gaussian take," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 17 346–17 356.

[3] S. Abdulwahab, H. A. Rashwan, N. Sharaf, S. Khalid, and D. Puig, "Deep monocular depth estimation based on content and contextual features," Sensors, vol. 23, no. 6, p. 2919, 2023.

[4] H. Lee, J. Park, W. Jeong, and S.-W. Jung, "Monocular depth estimation network with single-pixel depth guidance," Optics Letters, vol. 48, no. 3, pp. 594–597, 2023.

[5] Salim Aljawazneh, Q. S., & Ibrahim, H. (2019). Establishing technology for smart city development in Jordan's Amman-King Hussain Business Park. International Journal of Innovative Technology and Exploring Engineering, 8(5s), 213-220.

[6] Z.-F. Wang and Z.-G. Zheng, "A region based stereo matching algorithm using cooperative optimization," in 2008 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2008, pp. 1–8.

[7] N. Stefanoski, C. Bal, M. Lang, O. Wang, and A. Smolic, "Depth estimation and depth enhancement by diffusion of depth features," in 2013 IEEE International Conference on Image Processing. IEEE, 2013, pp. 1247–1251.

[8] D. Hoiem, A. N. Stein, A. A. Efros, and M. Hebert, "Recovering occlu- sion boundaries from a single image," in 2007 IEEE 11th International Conference on Computer Vision. IEEE, 2007, pp. 1–8.

[9] A. Saxena, M. Sun, and A. Y. Ng, "Make3d: Learning 3d scene structure from a single still image," IEEE transactions on pattern analysis and machine intelligence, vol. 31, no. 5, pp. 824–840, 2008.

[10] K. Karsch, C. Liu, and S. B. Kang, "Depth extraction from video using non-parametric sampling," in Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7- 13, 2012, Proceedings, Part V 12. Springer, 2012, pp. 775–788.

[11] J. Konrad, M. Wang, P. Ishwar, C. Wu, and D. Mukherjee, "Learning- based, automatic 2d-to-3d image and video conversion," IEEE Trans- actions on Image Processing, vol. 22, no. 9, pp. 3485–3496, 2013.

[12] M. K. Aicha Nouisser, Ramzi Zouari, "Deep learning based mobilenet and multi-head attention model for facial expression recognition," The International Arab Journal of Information Technology (IAJIT), vol. 20, no. 3, pp. 485 – 491, 2023.

[13] A. V. V. D. C. G. Surbhi Kapoor, Akashdeep Sharma, "A comparative study on deep learning and machine learning models for human action recognition in aerial videos," The International Arab Journal of Information Technology (IAJIT), vol. 20, no. 04, pp. 567 – 574, 2023.

[14] H. A. Owida, B. A.-h. Moh'd, N. Turab, J. Al-Nabulsi, and S. Abuowaida, "The evolution and reliability of machine learning techniques for oncology." International Journal of Online & Biomedical Engineering, vol. 19, no. 8, 2023.

[15] S. F. A. Abuowaida, H. Y. Chan, N. F. F. Alshdaifat, and L. Abualigah, "A novel instance segmentation algorithm based on improved deep learning algorithm for multi-object images," Jordanian J Comput Inf Technol (JJCIT), vol. 7, no. 01, pp. 10–5455, 2021.

[16] A. Torralba and A. Oliva, "Depth estimation from image structure," IEEE Transactions on pattern analysis and machine intelligence, vol. 24, no. 9, pp. 1226–1238, 2002.

[17] C. Chun, D. Park, W. Kim, and C. Kim, "Floor detection based depth estimation from a single indoor scene," in 2013 IEEE International Conference on Image Processing. IEEE, 2013, pp. 3358–3362.

[18] S. Choi, D. Min, B. Ham, Y. Kim, C. Oh, and K. Sohn, "Depth analogy: Data-driven approach for single image depth estimation using gradient samples," IEEE Transactions on Image Processing, vol. 24, no. 12, pp. 5953–5966, 2015.

[19] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," Advances in neural information processing systems, vol. 27, 2014.

[20] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab, "Deeper depth prediction with fully convolutional residual networks," in 2016 Fourth international conference on 3D vision (3DV). IEEE, 2016, pp. 239–248.

[21] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7132–7141.

[22] S. F. Abuowaida and H. Y. Chan, "Improved deep learning architecture for depth estimation from single image," Jordanian Journal of Computers and Information Technology, vol. 6, no. 4, 2020.

[23] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in Proceedings of the IEEE confer- ence on computer vision and pattern recognition, 2017, pp. 4700–4708.

[24] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from rgbd images," in European conference on computer vision. Springer, 2012, pp. 746–760.