ISPARTA UYGULAMALI BİLİMLER ÜNİVERSİTESİ

Uluslararası Teknolojik Bilimler Dergisi

International Journal of Technological Sciences

UTBD IJTS

Araştırma Makalesi/Research Article

# Monocular depth estimation and detection of near objects

**Ali Tezcan Sarızeybek***[1], **Ali Hakan Işık**[2]

[1]Burdur Mehmet Akif Ersoy University, The Graduate School of Natural and Applied Sciences, 15030, Burdur, Türkiye
[2]Burdur Mehmet Akif Ersoy University, Faculty of Engineering and Architecture, Department of Computer Engineering, 15030, Burdur, Türkiye

**Keywords**
Monocular Depth Estimation
Object Detection
Obstacle Avoidance
Image Processing

**Article history:**
Received: 20.09.2022
Accepted: 06.12.2022

**Abstract:** The image obtained from the cameras is 2D, so we cannot know how far the object is on the image. In order to detect objects only at a certain distance in a camera system, we need to convert the 2D image into 3D. Depth estimation is used to estimate distances to objects. It is the perception of the 2D image as 3D. Although different methods are used to implement this, the method to be applied in this experiment is to detect depth perception with a single camera. After obtaining the depth map, the obtained image will be filtered by objects in the near distance, the distant image will be closed, a new image will be run with the object detection model and object detection will be performed. The desired result in this experiment is, for projects with a low budget, instead of using dual camera or LIDAR methods, it is to ensure that a robot can detect obstacles that will come in front of it with only one camera. As a result, 8 FPS was obtained by running two models on the embedded device, and the loss value was obtained as 0.342 in the inference test performed on the new image, where only close objects were taken after the depth estimation.

**Atıf için/To Cite:** Sarızeybek A.T. Işık A.H. Monocular Depth Estimation and Detection of Near Objects. Uluslararası Teknolojik Bilimler Dergisi, 14(3), 124-131, 2022.

# Monoküler derinlik tahmini ve yakın nesnelerin tespiti

**Anahtar Kelimeler**
Monoküler Derinlik Tahmini
Nesne Tespiti
Engelden Kaçınma
Görüntü İşleme

**Makale geçmişi:**
Geliş Tarihi: 20.09.2022
Kabul Tarihi: 06.12.2022

**Öz:** Kameralardan elde edilen görüntü 2 boyutlu olduğu için cismin görüntü üzerinde ne kadar uzakta olduğunu bilemeyiz. Bir kamera sisteminde sadece belirli bir mesafedeki nesneleri algılamak için 2 boyutlu görüntüyü 3 boyutluya dönüştürmemiz gerekir. Derinlik tahmini, nesnelere olan mesafeleri tahmin etmek için kullanılır. 2 boyutlu görüntünün 3 boyutlu olarak algılanmasıdır. Bunu uygulamak için farklı yöntemler kullanılsa da, bu deneyde uygulanacak yöntem, tek bir kamera ile derinlik algısını tespit etmektir. Derinlik haritası elde edildikten sonra elde edilen görüntü yakın mesafedeki nesneler tarafından filtrelenecek, uzaktaki görüntü kapatılacak, nesne algılama modeli ile yeni bir görüntü çalıştırılacak ve nesne algılama gerçekleştirilecektir. Bu deneyde istenilen sonuç, düşük bütçeli projeler için çift kamera veya LIDAR yöntemlerini kullanmak yerine, bir robotun önüne gelecek engelleri tek kamera ile tespit edilmesini sağlamaktır. Sonuç olarak, gömülü cihaz üzerinde iki model çalıştırılarak 8 FPS elde edilmiş ve derinlik tahmini sonrası sadece yakın nesnelerin alındığı yeni görüntü üzerinde yapılan çıkarım testinde kayıp değeri 0.342 olarak elde edilmiştir.

## 1. Introduction

Depth perception is a big problem for most technologies in the world. Depth perception is used from robotic applications to mobile devices, generally two cameras that can measure depth by combining the images taken from the right and left camera are used for depth detection [1]. Depth estimation, for example, in smart cleaning robots produced in the robotic field, is programmed to determine where the robot is and where the robot cannot enter, and to decide where the robot should clean according to the distance of the objects [2]. In mobile devices, depth perception is used to detect the close object and add a blur effect on other objects so that the Bokeh Effect, called portrait mode, can be applied on the image [3, 4, 5]. In this section,

* İlgili yazar/Corresponding author: atsarizeybek@mehmetakif.edu.tr

depth perception, various usage methods and the developed application will be discussed.

Dual cameras are a problem in terms of portability in mobile devices and cost and size in robotic applications. In SLAM applications where a single camera is used, cameras that can detect infrared rays and measure distances according to the return time of the rays are also used [6, 7], but this method is also very costly and looks bad in terms of aesthetics. The LIDAR method can instantly obtain the distances of the surrounding objects. 360 degree rotating LIDAR technology is able to obtain depth values of the entire environment [8]. LIDAR, which is generally used in robots, can calculate the depth of surfaces that are equivalent to their height. The linear SLAM used in mobile phones calculates the depth in the direction of the camera [9]. Since LIDAR technology is high in terms of cost, it cannot be used in works that are planned to be low-cost.

A system that can perform monocular depth estimation is recommended so that depth perception can be used in projects with low cost or designed to have a single camera in the design, where only one camera is needed. In the proposed system, depth perception will be realized with a single camera and it will be fast in terms of speed performance. In the robotic application developed to test the system, the robot will stop when it sees a living object in front of it, and give a warning when it sees a non-living object. The reason for the development of this application is that in the feed pushing robot used in animal husbandry, a camera is placed on the robot in order to stop or pass by an object in case of forward movement, but since the camera is kept facing the opposite direction, it will detect objects at a distance, so it will only detect the most distant objects with depth estimation. Detection of nearby objects is desired. In the robotic application developed using depth perception, the depth will be estimated using the monocular depth estimation model, and after a certain depth on the image, it will be removed from the image.

### *1.1. Stereo Depth Estimation*

With the dual camera, depth perception is applied based on the optimal fit pixels of the epipolar lines in the right and left images [10]. Depth estimation is performed by comparing the common pixels of the right camera and the left camera [11]. For point matching, features are extracted on the image and a feature map of their behavior at different inequality levels is drawn. Using the inequality calculation formula, the volume is calculated from the feature map, and if the performance is unsatisfactory, the inequality is refined [12, 13]. Figure 1 shows the basic logic of stereo depth estimation.

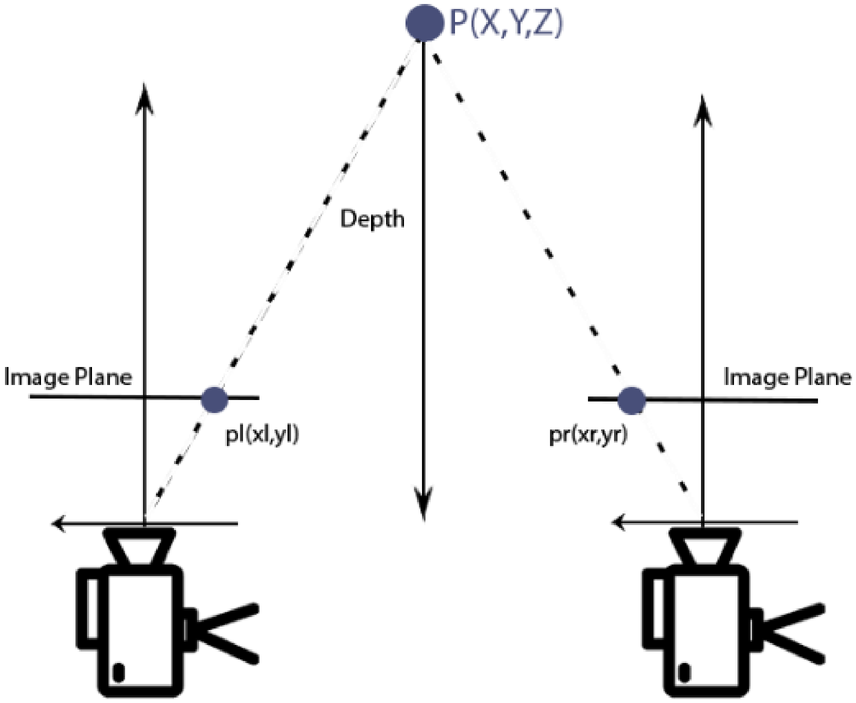

Fig. 1. Logic of stereo depth estimation, depth perception is made by calculating the intersection of the image plane of the two cameras. [14]

### *1.2. Monocular Depth Estimation*

Monocular depth estimation is a single-shot depth estimation method [15]. After the model is trained with a data set prepared on various images, depth map of images and masking data, the depth estimation model is obtained. KITTI and NYUV2 datasets are the most used depth estimation datasets [16]. It is used for autonomous driving [17], 3D modeling [18] and augmented reality [19]. In depth estimation models, performance comparison is measured by loss values [20]. Loss values are calculated on different data sets and models are compared accordingly. Usually, loss values are calculated by combining two data sets [21]. Depth estimation can be used to measure the surface of an object, calculate the distance between two objects, and map the desired location. MiDas, Binsformer, GCNDepth, LeReS, RPSF, DepthFormer monocular depth estimation methods are used for depth sensing.

## 2. Related Works

Miangoleh et al. (2021), it was aimed to eliminate the inadequacy of the monocular depth estimation model in the proposed method. The shortcomings of the model are depth maps at sub megapixel resolution, and lack of fine detail in inferences. In the proposed method, the binary estimation method is presented, and as a result, a method that can obtain high quality depth maps has been developed. [22]

Li et al. (2021), a new metric was proposed for monocular depth estimation. As a result of the tests conducted with the NYU Depth v2 data set, it was determined that the depth estimation process over video, together with the suggested metrics, was more performant in terms of intensity. [23]

Jung et al. (2021), it was aimed to make depth estimation easier on images with moving objects. A learning based method called DnD has been developed to estimate density depth maps, and monocular depth estimation is performed after the image is processed with depth, RGB encoder and decoder by combining SfM and MVS algorithms. [24]

Kopf et al. (2021), aims to obtain consistent density depth maps over images. An optimization algorithm has been proposed for this purpose, the DeepV2D method has been compared by using methods such as depth filter, flexible pose and depth fine tuning, and it has been concluded that it is more efficient in terms of performance. [25]

In the study by Chang and Werzstein (2019), a deep optics paradigm was proposed for 3D object detection and depth estimation. Optimization schemes and coding strategies were created using NYUv2, KITTI and Rectangles datasets. As a result, it was stated that chromatic aberrations affect the depth estimation results well. Object detection model is trained in KITTI dataset, improved 3D object detection is provided for depth estimation. [26]

As understood from the literature summary, new monocular depth estimation models were created by combining Monocular Depth Estimation and monocular depth estimation data sets and adding various filters on the image data, or performance tests of the models created by adding new performance metrics were performed and comparisons were made between the methods. In the study, the performance of running the MiDaS monocular depth estimation model simultaneously with the object detection model in an embedded system was measured.

## 3. Method and Material

### *3.1. MiDaS Monocular Depth Estimation*

MiDaS, one of the monocular depth estimation models was used in the study. Monocular depth estimation is used to make depth estimation with a single camera. MiDaS datasets contain original image, depth values and masked images [27]. The architecture of the MiDaS model is based on the ResNet architecture. For model training, 10 data sets, including HRWSI, TartanAir, IRS, ReDWeb, DIML, Movies, MegaDepth, WSVD, ApolloScape and BlendedMVS, were used. For higher performance values, images of 23 3D movies are included in the dataset. The model trained on datasets was developed with multi-objective optimization. MiDaS v2.1 Small model, which has the highest performance among small MiDaS models, was used. [28]

The performance values of the models are given in Table 1.

As seen in the table, the best model in terms of performance with the lower the better approach is the MiDaS v2 small model. Although the performance is higher in v2, v2.1 is much better in terms of FPS. Therefore, MiDaS v2.1 small model was preferred as the model.

Table 1. Performances of MiDaS models (The lower is better) [28]

| Model | DIW,WHDR | Eth3d, AbsRel | Sintel, AbsRel | TUM | Kitti | NyuDepth | FPS |
|---|---|---|---|---|---|---|---|
| MiDaS v2 Small | **0.1248** | 0.1550 | **0.33** | 17 | **21.81** | 15.73 | 0.6 |
| **MiDaS v2.1 Small** | 0.1344 | **0.1344** | 0.337 | **14.53** | 29.27 | **13.43** | **30** |

### *3.2. Object Detection*

Object detection is an image processing technology based on Convolutional neural networks [29]. Labeling the images and then training the features for that class within the tags in the image is essential for model training. For example, there are studies such as following the ball or players in sports competitions [30] and lane tracking in unmanned vehicles [31]. In this study, SSD ResNet50 model, which is one of the mobile object detection models, was used. The COCO mAP value of the model was 35 and the speed was 76ms, so it was used. TFLite which can easily work on mobile and embedded devices was used as a toolkit [32].

### *3.3. Feed Pushing Robot*

The feed pushing robot is produced to push the feed towards the cows at the desired time and route. In the project, there is a camera on the feed pushing robot. The reason for the existence of this camera is the safety precautions that must be taken in case of living or inanimate objects in the direction of the robot while pushing the feeds with the helix.

### *3.4. Nvidia Jetson Nano*

Nvidia Jetson Nano has been added to the robot for the image processing stages, and the image processing steps are completely handled by Jetson Nano. Although it is difficult for two models to run, performance around 8 FPS is achieved. This is a sufficient value for safety controls according to the slow speed of the robot.

## 4. Discussion and Results

In this study, The application aims at object detection with monocular depth estimation. The camera on the robot is looking in the direction the robot will go, so since the camera can detect the object even 50 meters away, it is aimed to detect only objects at a certain distance with the depth estimation and the robot to stand accordingly. In this way, if an object comes across while moving, the robot will be able to stop automatically. In Table 2, the frame per-second values obtained as a result of the tests performed on Nvidia Jetson Nano, as a result of running the Object Detection and Depth Estimation models separately and together are given.

Table 2. FPS results of two models,

| **Model** | **FPS** |
|---|---|
| Object Detection | 25 |
| Depth Estimation | 28 |
| Object Detection + Depth Estimation | 8 |

After the depth estimation and object detection models were combined, they were tested on the validation data set in the object detection model, and 0.342 loss was obtained in the categorical cross entropy function.

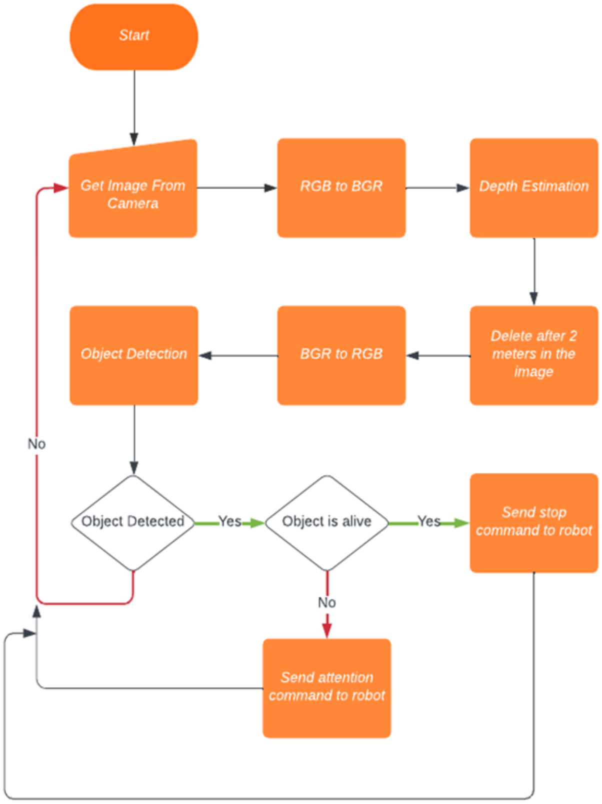

Figure 2. Explaining the logic of the application as a diagram

As seen in the diagram, an instant image is taken from the camera, the RGB image is converted to BGR because the depth estimation model requests a BGR image. After the depth mapping, the area corresponding to approximately 2 meters on the image is subtracted from the image. The resulting new image is then converted back from BGR to RGB and given to the object detection model. If an object is recognized as output, it is checked whether the object is alive. If the object is not detected or the processes are finished, it returns to the image acquisition process from the camera.

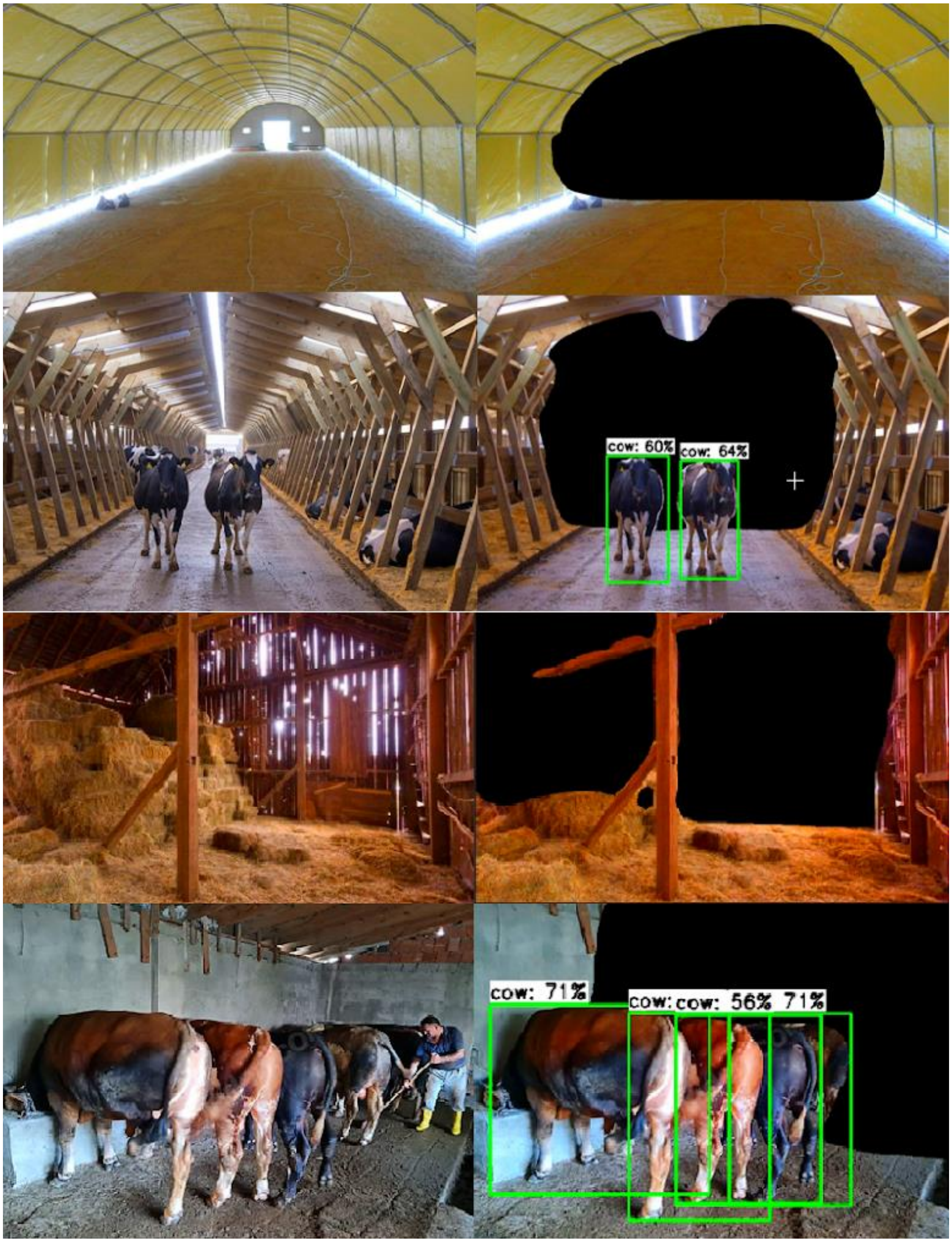

Figure 3. Images with depth estimation, in which only close objects are obtained, with object detection on it. On the left are the original images, on the right are the images with depth perception and object detection applied in the application.

As can be seen in Figure 3, there are original versions of the images on the left. After the image is given to the depth perception model, only the image obtained after the acquisition of close objects is given to the object detection model. The Object detection model also detects objects on the new image.

## 5. Conclusion and Future Work

The aim of this study is to detect nearby objects and to use this study in the feed pushing robot used in the robotic field. After the depth detection is done, after deleting the image after a certain interval on the image, the image is given as input to the object detection model, and the objects recognized on the image and their coordinates as pixels are obtained as output. Although the subject of using two different models has been mentioned a lot in the literature, this subject has not been mentioned in the field of animal husbandry robotics. Although the use of this work in the feed pushing robot is very important for safety, operations such as lane tracking, cow detection and adjusting the feed pushing according to the position of the cow can be performed in future studies.

## References

[1] Kusupati, U., Cheng, S., Chen, R., & Su, H. (2020). Normal assisted stereo depth estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 2189-2199).

[2] Hess, J., Beinhofer, M., & Burgard, W. (2014, May). A probabilistic approach to high-confidence cleaning guarantees for low-cost cleaning robots. In 2014 IEEE international conference on robotics and automation (ICRA) (pp. 5600-5605). IEEE.

[3] Wang, Y., Lai, Z., Huang, G., Wang, B. H., Van Der Maaten, L., Campbell, M., & Weinberger, K. Q. (2019, May). Anytime stereo image depth estimation on mobile devices. In 2019 international conference on robotics and automation (ICRA) (pp. 5893-5900). IEEE.

[4] Dutta, S., Das, S. D., Shah, N. A., & Tiwari, A. K. (2021). Stacked Deep Multi-Scale Hierarchical Network for Fast Bokeh Effect Rendering from a Single Image. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 2398-2407).

[5] Ignatov, A., Malivenko, G., Plowman, D., Shukla, S., & Timofte, R. (2021). Fast and accurate single-image depth estimation on mobile devices, mobile ai 2021 challenge: Report. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 2545-2557).

[6] Collis, R. T. H. (1969). Lidar. In Advances in Geophysics (Vol. 13, pp. 113-139). Elsevier.

[7] Hecht, J. (2018). Lidar for self-driving cars. Optics and Photonics News, 29(1), 26-33.

[8] Wróżyński, R., Pyszny, K., & Sojka, M. (2020). Quantitative landscape assessment using LiDAR and rendered 360 panoramic images. Remote Sensing, 12(3), 386.

[9] Ullrich, A., & Pfennigbauer, M. (2016, May). Linear LIDAR versus Geiger-mode LIDAR: impact on data properties and data quality. In Laser Radar Technology and Applications XXI (Vol. 9832, pp. 29-45). SPIE.

[10] Long, X., Liu, L., Li, W., Theobalt, C., & Wang, W. (2021). Multi-view depth estimation using epipolar spatio-temporal networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 8258-8267).

[11] Kusupati, U., Cheng, S., Chen, R., & Su, H. (2020). Normal assisted stereo depth estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 2189-2199).

[12] Ding, X., Xu, L., Wang, H., Wang, X., & Lv, G. (2011). Stereo depth estimation under different camera calibration and alignment errors. Applied Optics, 50(10), 1289-1301.

[13] Wang, Y., Lai, Z., Huang, G., Wang, B. H., Van Der Maaten, L., Campbell, M., & Weinberger, K. Q. (2019, May). Anytime stereo image depth estimation on mobile devices. In 2019 international conference on robotics and automation (ICRA) (pp. 5893-5900). IEEE.

[14] Fahmy, A. A., Ismail, O., & Al-Janabi, A. K. (2013). Stereo vision based depth estimation algorithm in uncalibrated rectification. Int J Video Image Process Netw Secur, 13(2), 1-8.

[15] 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(9), 1612-1627.

[16] Yuan, W., Gu, X., Dai, Z., Zhu, S., & Tan, P. (2022). NeW CRFs: Neural Window Fully-connected CRFs for Monocular Depth Estimation. arXiv preprint arXiv:2203.01502.

[17] Xue, F., Zhuo, G., Huang, Z., Fu, W., Wu, Z., & Ang, M. H. (2020). Toward hierarchical self-supervised monocular absolute depth estimation for autonomous driving applications. In 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (pp. 2330-2337). IEEE.

[18] Huynh, L., Nguyen-Ha, P., Matas, J., Rahtu, E., & Heikkilä, J. (2020, August). Guiding monocular depth estimation using depth-attention volume. In European Conference on Computer Vision (pp. 581-597). Springer, Cham.

[19] Ramamonjisoa, M., & Lepetit, V. (2019). Sharpnet: Fast and accurate recovery of occluding contours in monocular depth estimation. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (pp. 0-0).

[20] Lee, J. H., & Kim, C. S. (2020, August). Multi-loss rebalancing algorithm for monocular depth estimation. In European Conference on Computer Vision (pp. 785-801). Springer, Cham.

[21] Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., & Koltun, V. (2020). Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. IEEE transactions on pattern analysis and machine intelligence.

[22] Miangoleh, S. M. H., Dille, S., Mai, L., Paris, S., & Aksoy, Y. (2021). Boosting monocular depth

estimation models to high-resolution via content-adaptive multi-resolution merging. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 9685-9694).

[23] Li, S., Luo, Y., Zhu, Y., Zhao, X., Li, Y., & Shan, Y. (2021). Enforcing Temporal Consistency in Video Depth Estimation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 1145-1154).

[24] Jung, D., Choi, J., Lee, Y., Kim, D., Kim, C., Manocha, D., & Lee, D. (2021). DnD: Dense Depth Estimation in Crowded Dynamic Indoor Scenes. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 12797-12807).

[25] Kopf, J., Rong, X., & Huang, J. B. (2021). Robust consistent video depth estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 1611-1621).

[26] Chang, J., & Wetzstein, G. (2019). Deep optics for monocular depth estimation and 3d object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 10193-10202).

[27] Gkikas, A., Proestakis, E., Amiridis, V., Kazadzis, S., Di Tomaso, E., Marinou, E., ... & García-Pando, C. P. (2022). Quantification of the dust optical depth across spatiotemporal scales with the MIDAS global dataset (2003–2017). Atmospheric Chemistry and Physics, 22(5), 3553-3578.

[28] Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., & Koltun, V. (2020). Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. IEEE transactions on pattern analysis and machine intelligence.

[29] Zhiqiang, W., & Jun, L. (2017, July). A review of object detection based on convolutional neural network. In 2017 36th Chinese control conference (CCC) (pp. 11104-11109). IEEE.

[30] Zhou, X., Gong, W., Fu, W., & Du, F. (2017, May). Application of deep learning in object detection. In 2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS) (pp. 631-634). IEEE.

[31] Kemsaram, N., Das, A., & Dubbelman, G. (2019, July). An integrated framework for autonomous driving: object detection, lane detection, and free space detection. In 2019 Third World Conference on Smart Trends in Systems Security and Sustainability (WorldS4) (pp. 260-265). IEEE.

[32] Black, A. W., & Lenzo, K. A. (2001). Flite: a small fast run-time synthesis engine. In 4th ISCA Tutorial and Research Workshop (ITRW) on Speech Synthesis.

*Bu çalışma 20-22 Mayıs 2022 tarihinde, Bakü, Azerbaycan'da ICAIAME 2022 (International Conference on Artificial Intelligence and Applied Mathematics in Engineering) konferansında sunulmuştur.*