Deep Neural Networks for Accurate Depth Estimation with Latent Space Features

Hyunsik Ahn * and Siddiqui Muhammad Yasir

Posted Date: 12 November 2024

doi: 10.20944/preprints202411.0868.v1

Keywords: visual learning; deep learning in robotics and automation; computer vision for automation; depth estimation; monocular depth estimation; multi-scope vision

 doi:10.20944/preprints202411.0868.v1

Article

Deep Neural Networks for Accurate Depth Estimation with Latent Space Features

Siddiqui Muhammad Yasir [1] and Hyunsik Ahn [2,*]

[1] Department of Mechanical System Engineering, Tongmyong University, Busan, South Korea
[2] School of Artificial Intelligence, Tongmyong University, Busan, South Korea
* Correspondence: hsahn@tu.ac.kr

Abstract: Depth estimation plays a pivotal role in advancing human-robot interactions, especially in indoor environments where accurate 3D scene reconstruction is essential for tasks like navigation and object handling. Monocular depth estimation, which relies on a single RGB camera, offers a more affordable solution compared to traditional methods that use stereo cameras or LiDAR. However, despite recent progress, many monocular approaches struggle with accurately defining depth boundaries, leading to less precise reconstructions. In response to these challenges, this study introduces a novel depth estimation framework that leverages latent space features within a deep convolutional neural network to enhance the precision of monocular depth maps. The proposed model features dual encoder-decoder architecture, enabling both color-to-depth and depth-to-depth transformations. This structure allows for refined depth estimation through latent space encoding. To further improve the accuracy of depth boundaries and local features, a new loss function is introduced. This function combines latent loss with gradient loss, helping the model maintain the integrity of depth boundaries. The framework is thoroughly tested using the NYU Depth V2 dataset, where it sets a new benchmark, particularly excelling in complex indoor scenarios. The results clearly show that this approach effectively reduces depth ambiguities and blurring, making it a promising solution for applications in human-robot interaction and 3D scene reconstruction.

Keywords: visual learning; deep learning in robotics and automation; computer vision for automation; depth estimation; multi-scope vision

1. Introduction

In human-robot interaction, comprehending the spatial relationships of 3-dimensional (3D) surroundings is a critical perceptual undertaking for various robotic applications, encompassing manipulation, exploration, and navigation [1]. Accurate depth perception is typically required by robots to evade obstructions and handle objects. In industrial settings, moving agents, i.e. autonomous ground vehicles (AGV), robot arms often possess a color camera for surveillance purposes. However, depth estimation frequently necessitates specialized equipment, such as stereo cameras, structured-light sensors, or time-of-flight sensors, which are relatively costly compared to a single RGB camera. While researchers have made strides in monocular depth perception, there remains significant room for improvement [2].

In recent times, there has been a growing interest in a technique used to deduce depth information from a single monocular image. This depth information is valuable as it provides useful clues for the efficient understanding of a given scene such as the vanishing point location and horizontal boundary. Due to this, various computer vision applications now consider depth estimation as a crucial prerequisite for sophisticated operations, including 3D scene modeling & reconstruction. Depth estimation has become vital for autonomous driving systems as it enables the interpretation of the geometric structure in acquired images.

The stereo-matching principle is a fundamental technique for estimating depth using multiple cameras. The stereo sensor comprises two horizontally displaced cameras, enabling the corresponding pixels in both cameras to be aligned on the same horizontal line. Consequently, stereo matching can generate a disparity map that depicts the positional differences between corresponding

 doi:10.20944/preprints202411.0868.v1

pixels in stereo images [3]. In contrast, structure from motion (SFM) and multi-view stereo (MVS) do not limit the camera poses, causing the pixel correspondence to be non-linear and making it more challenging to establish pixel correspondences [4,5]. The concept of applying stereo matching multi-scope vision is centered on acquiring high-quality depth maps with a single camera and controlled motion, whereby more constraints can be imposed on the reconstructed depth maps. When multiple images captured at aligned camera positions are used for depth estimation, it is referred to as the multi-scope vision, analogous to stereo vision using two horizontally aligned images. Motivated by the principle that depth estimation with two perfectly aligned images in stereo vision is relatively simpler than using two images with unknown camera poses, this paper contends that capturing multiple images with aligned single camera positions could lead to more precise and robust depth estimation.

Although there have been great strides in estimating depth from stereo images and videos, monocular depth estimation is still difficult because of the uncertainty caused by the ill-posed nature of the problem [6]. To address this limitation, statistical feature-based approaches were initially studied. These approaches, represented by methods in categories such as graph optimization and pixel clustering, typically begin by conducting image segmentation. The original distributions of extracted features from individually segmented region, such as edge orientations, textures, frequency coefficients, etc., are then aggregated to learn the corresponding depth values [7]. Other approaches have attempted to assign perceptually appropriate depth values to an image based on its structural similarity with other scenes [8]. Although statistical feature-based methods have shown promise in inferring depth information from a single monocular image, they have limitations in accommodating the diverse variations of geometric structures, especially in complicated indoor environments. Recently, several researchers have started to apply the generative model via deep neural networks for estimating depth values from a single monocular image, inspired by its success in various other fields. The depth estimation problem is reformulated as the image-to-image generation problem, where the input color image is converted into a corresponding depth image. The geometric features that are virtuous at revealing the depth structures of a given scene are learned efficiently through deep-layered architectures, without the need for designing hand-crafted features. The convolutional neural network (CNN) with variable receptive fields is popularly employed to support the learning process between color and depth images. Large-sized datasets constructed using Radar or LiDAR-based capturing systems, such as the NYU v2 dataset [9], are adopted to train deep neural networks effectively.

This paper proposes a lightweight human-robot interaction system for depth estimation from a single RGB image that efficiently guides the learning process of the RGB image-to-depth relationship by exploiting features extracted from the latent space network. These encoded features contain the geometrical structure compactly, which is relevant to the depth layout of the given scene, and thus the corresponding gradients sharpen the depth boundary efficiently. By considering the inherent properties of the depth generation and the relationship between color and depth values, the proposed method can reduce depth ambiguity in homogeneous regions and blurring artifacts at depth boundaries. This approach is, a biomimetic approach, ampesizing low level visual perception using skip connection, which is similar to stressing basic visual processing of primary visual cortex for visual understading of brains. The main contributions of the proposed method are summarized as follows:

- We propose a human-robot interaction system for a monocular depth estimation auto-encoder network to effectively learn the complex process of transforming a color image into a depth image. Unlike previous approaches that relied on the concept of perceptual loss.
- The proposed technique aims to learn the process of “generation” from the latent space rather than using a “classification” strategy to refine the estimated depth information.
- In contrast to other techniques, the proposed method works reasonably accurately since the proposed network design does not use feature branches other than skip connections of residula blocks.

This paper's summary is organized as follows. In Section 2, a comparison of comparable studies is reviewed. Section 3 provides a thorough explanation of the proposed human-robot interaction system for monocular depth estimation utilizing human-robot interaction to perceive an indoor

 doi:10.20944/preprints202411.0868.v1

environment. In Section 4, experimental findings are illustrated using a benchmark dataset. In Section 5, the results and conclusion are presented.

2. Related Work

In this section, we first go through previous systems for capturing images for depth estimation and then explain algorithms that combine data to create depth maps for service robot systems.

Fang et. al., [10] propose a performance criterion for the depth estimation of an active vision system. López-Nicolás et. al., [11] present a new visual servoing approach for mobile robots with a fixed monocular system on board. Sabnis et. al., [12] present a depth estimation technique based on the defocus blur associated with a camera setting. Turan et. al., [13] present monocular endoscopic capsule robots that can measure depth and odometry in real time without supervision. Jin et al. [14] conduct specific behavioral tasks and autonomously improve a humanoid robot's depth-estimation accuracy, with a novel PA-based cyclic learning framework. By mimicking the human finger touch Xiao et. al., [15] propose a tactile sensing-based deep recurrent neural network (DRNN) with long short-term memory (LSTM) architecture to improve the accuracy of the detection and depth estimation of tumors embedded in soft tissue. The proposal of Cheng et. al., [16] is to build a modular interactive framework based on RGB images, which aims to improve the phenomena of high dependence on depth sensor camera and low adaptability to distance in human-robot interaction (HRI) framework by using advanced human pose estimation technology. A disparity estimation neural network for an electric inspection robot is proposed, which consists of two main parts: PSMNet module and lightweight cutting module Yu et. al., [17]. Other influential work includes Wang et. al., [18] Describe a novel method for estimating the scene structure and stiff body motion characteristics from monocular image sequences. With a monocular image sequence devoid of depth data, Shimada et al., [19] propose a method to properly estimate the posture e.g., joint angles of a moving human hand as well as to fine-tune the hand model's 3D shape i.e., widths & lengths. Based on monocular vision imaging, a high accuracy and efficiency estimation algorithm of the relative pose of cooperative space targets is presented. To increase the accuracy of feature extraction and estimation efficiency, multiple target tracking techniques are used, while the Levenberg-Marquardt method (LMM) is used to achieve a well-global convergence Pan et. al., [20]. Gysel et. al., [21] introduce a unique latent vector space model that, without the need for explicit annotations, jointly learns the latent representations by marginalizing depth-map changes and outliers, Kashyap et al., [22] introduce a learning-based technique to directly estimate camera motion parameters from optic flow. To quantify the discrepancies between the prediction and ground truth in the hierarchical embedding spaces of depth maps, Wang et al., [18] present a hierarchical loss for monocular depth estimation. Reading et. al., [23] Develop CaDDN as a completely differentiable, end-to-end method for detecting objects and estimating depth simultaneously. Proposed a Multi-Scale Features Network (MSFNet) that comprises an Enhanced Diverse Attention (EDA) module and an Up-Sample-Stage Fusion (USF) module for better depth estimation Pei, [24]. The aforementioned unsupervised framework cannot be directly applied to some difficult conditions, such as night and rainy nights, where the key photometric consistency hypothesis is impractical due to the complicated illumination and scenarios. Zhao et. al., [25] Examine the issue of unsupervised monocular depth estimation in extremely complicated settings, and use an image transfer-based domain adaption approach to tackle this difficult issue in human-robot interaction (HRI). Reducing model computational complexity while retaining high accuracy performance is the goal of Guo et al., [26].

Despite the advancements made by generative models employing deep neural networks in estimating depth information from monocular images, existing approaches continue to struggle with clearly revealing depth boundaries, resulting in blurry restoration results. This paper presents a new, straightforward method for depth estimation from a single image that addresses the issue of blurring artifacts at the depth edges. Technical specifics will be discussed in the subsequent section.

 doi:10.20944/preprints202411.0868.v1

3. Proposed Monocular Depth Estimation

This chapter details the proposed monocular depth estimation model and its training methodology. The ultimate task of this approach is to acquire better depth information to reconstruct a 3D model of the environment from which the robot can understand the surroundings and serve humans. The aim of this research is to explore the generative process that creates the depth arrangement from the geometric structure of a monochromatic image. In order to maintain the depth boundary, we utilize a deep convolutional neural network to compactly encode the single RGB to depth relationship in the latent space. This helps us efficiently enhance the quality of the depth map obtained from the color input image in our proposed method. In order to accomplish this, we will initially present the general structure of our two deep neural networks. Following that, we will provide a detailed explanation of the complete depth estimation procedure utilizing our training approach. Ultimately, a detailed explanation is provided for the loss function used, which includes data loss, latent loss, and gradient loss.

3.1. Depth Estimation Deep Learning Model

The proposed architecture for depth estimation consists of two encoder-decoder networks: depth-to-depth and color-to-depth networks. Both networks have a similar structure consisting of three main elements: encoder, ResBlocks, and decoder. The general layout is depicted in Figure 1. The input image is compressed into latent features effectively by multiple ResBlocks on the encoder side, with a slightly modified version from the original residual network shown in Figure 2.

As is commonly known in brain science, the primary visual cortex (V1) detects basic visual characteristics such as shape, color, contrast ratio, and line direction, and the secondary visual cortex (V2) uses the detection results of the V1 to recognize a higher level of visual perception such as depth and relationships between objects. Therefore, for detecting the edge of depth more clearly, the role of V1 percepting the primary visual information is important. The skip connection of the residual block in this approach has the effect of highlighting the primary visual information of image. By layering a sufficient number of ResBlocks, latent features implicitly encode characteristics for generating depth. The small spatial size of these highly encoded latent features holds necessary information to reconstruct the target image, specifically the depth map. Batch normalization and ReLU layers follow every convolution layer except the last output layer. On the decoder side, the feature map size is doubled through up-sampling using bilinear interpolation. The depth map is effectively produced by the symmetric decoder using the latent features.

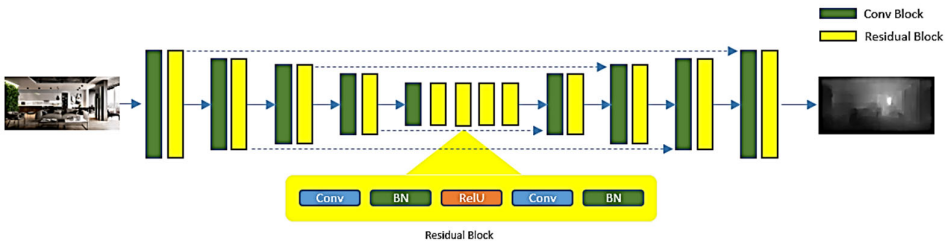


Figure 1. The general structure of the method suggested for estimating depth. The depth generation network being suggested consists of both convolution and deconvolution layers. In order to effectively understand how color and depth are related in each image, the suggested network is trained with a loss function that includes features from the latent space of the network.

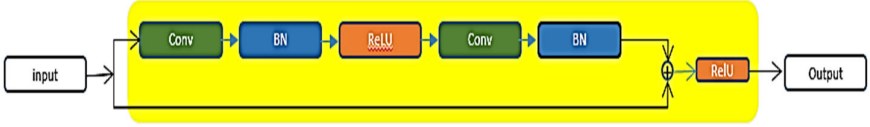


Figure 2. Detailed view of Residual Block.

 doi:10.20944/preprints202411.0868.v1

Every level of feature maps effectively captures the data that signifies the internal correlation within each spatial dimension. Skip connections, as described in [27], are utilized in the color-to-depth network for bringing back local details. The suggested approach can effectively capture both specific elements and overall structures of the depth map by utilizing learned latent features related to depth generation. These underlying characteristics influence the final outcome, the depth-map created from the initial color image, to resemble the real depth map. Because of the implicit enhancement provided by this guided network, the proposed scheme is able to detect the depth boundary in the estimated result, even in complex outdoor settings.

The comprehensive structure of the network as proposed can be found in Figure 2. The sizes of the convolution filters vary from 3x3 to 9x9 based on the convolution layers they are used in. By utilizing the latent features learned during the depth generation process, the suggested network can reconstruct both local details and global layouts of the depth map. A thorough description of the training approach for these two networks will be given in the following subsection.

Table 1. Detailed architecture of the proposed depth estimation network.

Module	Layer type	Weight dimension	Stride
Encoder	Conv	64x3x9x9	1
	ResBlock	64 x 64 x 9 x 9	1
	Conv	128 x 64 x 7 x 7	2
	ResBlock	128 x 128 x 7 x 7	1
	Conv	256 x 128 x 5 x 5	2
	ResBlock	256 x 256 x 5 x 5	1
	Conv	512 x 256 x 3 x 3	2
	ResBlock	512 x 512 x 3 x 3	1
	Conv	512 x 512 x 3 x 3	2
ResNet	6x ResBlock	512 x 512 x 3 x 3	1
Decoder	Upsampling	-	-
	Conv	512 x 512 x 3 x 3	1
	ResBlock	512 x 512 x 3 x 3	1
	Upsampling	-	-
	Conv	256 x 512 x 3 x 3	1
	ResBlock	256 x 256 x 5 x 5	1
	Upsampling	-	-
	Conv	256 x 128 x 5 x 5	1
	ResBlock	128 x 128 x 7 x 7	1
	Upsampling	-	-
	Conv	128 x 64 x 7 x 7	1
	ResBlock	64 x 64 x 9 x 9	1
	Conv	64x3x9x9	1

3.2. Latent Loss Functions

Once the predicted depth map is obtained from the color-to-depth network R and the corresponding ground truth, the guided network G is used to extract dense features from each ResBlock on the encoder side. In order to determine the loss value, we adopt a comparable method to Johnson et al. [28] that involves measuring the disparity between features extracted from the activation layer of each scale and the latent space. Our proposed approach involves using the guided network G for depth restoration in a depth-to-depth auto-encoder, instead of relying on pre-trained models like VGG, ResNet, etc. This method focuses on teaching the decoder the process of 'generation' from the latent space, rather than using a 'classification' strategy in perceptual loss:

$$L_l(G(y), G(y^*)) = \sum_j \left(\frac{1}{N_j} \sum_k^{N_j} || G_j(y_k) - G_j(y_k^*) || \frac{2}{2} \right) \quad (1)$$

 doi:10.20944/preprints202411.0868.v1

The latent loss function is determined by adding up the discrepancy between $G_j(y)$ and $G_j(y*)$, are feature maps originating from the activation layer preceding the j[th] convolution layer. $G_j(yk)$ and $G_j(y*k)$ represent the kth feature value extracted from the estimated depth map y and the ground truth y∗k, respectively. N_j represents the dimension of the feature map obtained from the j[th] activation layer. The aim of the proposed latent loss function is to make the depth map estimated from the latent space of the top encoded layer equal to the ground truth at the feature level.

3.3. Gradiant Loss

In order to, improve the details effectively at depth boundaries, a crucial issue to address, we suggest utilizing both image-level and feature level gradients in the loss function. More precisely, the consistency of gradients between the predicted depth map and its corresponding ground truth (at the image level) is calculated in the following manner:

$$L_{gd}(y, y^*) = \frac{1}{N}\sum_{i}^{N} |y_{h,i} - y^*_{h,i}| + |y_{v,i} - y^*_{v,i}| \quad (2)$$

where $y_{h,i}$ and $y_{v,i}$ represent the i^{th} gradient value of the estimated depth map in the horizontal and vertical directions, respectively. In the same way, $y^*_{h,i}$ and $y^*_{v,i}$ represent the gradient value of the actual data in both directions. The depth map consists of a total of N pixels. Furthermore, the gradient loss is determined by calculating the gradient of encoded features collectively:

$$L_{gl}(G(y), G(y^*)) = \sum_{j} \left(\frac{1}{N_j} \sum_{k}^{N_j} | G_{h,j}(y_k) - G_{h,j}(y^*_k)| + |G_{v,j}(y_k) - G_{v,j}(y^*_k)| \right) \quad (3)$$

where $G_{h,j}$ & $G_{v,j}$ represents gradient of encoded features of specific inputs in the horizontal & vertical directions, respectively. By incorporating both terms into the ultimate loss function, the depth map's high-frequency components can be effectively enhanced. A key benefit is that the encoded-features within proposed network effectively represent the depth structure in a condensed manner across multiple scales, leading to significant assistance from their gradients in enhancing the clarity of depth boundaries as illustrated in Figure 4. Hence, the gradient of the encoded-features suggested in this article is believed to be helpful in effectively recovering the depth boundary.

To summarize, extracted features from the guided network's latent space enable learning the intricate color-depth relationship in the proposed method (refer to (1) and (2)). The depth layout's core structure can be accurately reconstructed from just one monocular image due to its inherent properties of depth generation being condensed. Additionally, the gradients of these encoded features have shown to be effective in recovering the depth boundary, a task that has proven challenging for previous techniques.

4. Experiment

We trained our model using the original NYU Depth v2 [9] dataset collected indoors. The unprocessed datasets have many extra pictures gathered from identical locations as those in the popular smaller datasets, but with no prior editing. More precisely, areas that do not have a depth measurement are left blank. However, our model is inherently equipped to deal with such gaps, and its need for a substantial training set makes these original distributions valuable sources of data.

4.1. NYU v2 Depth Dataset

The dataset known as NYU Depth [9] is comprised of 464 indoor scenes that were recorded as videos. The official train & test division with 249 scenes for training and 215 for testing is utilized, and the training set is created using the raw data from these scenes. The RGB inputs are reduced by half in size from 640x480 to 320x240. As the depth and RGB cameras have varying frame rates, each depth image is matched with the closest RGB image in time, resulting in discarding frames where one RGB image corresponds to multiple depth images. The dataset's camera projections are utilized to align RGB and depth pairs, with pixels lacking depth values being excluded. Invalid areas due to

 doi:10.20944/preprints202411.0868.v1

windows and shiny surfaces are eliminated by masking depths that match the minimum or maximum values in each image. A pair of images is shown in Figure 3, with the training set having 120,000 distinct images rearranged into a list of 220,000 by evening out the scene distribution (1200 images per scene). The model undergoes testing on the 694-image NYU Depth v2 test set (with depth values filled in).

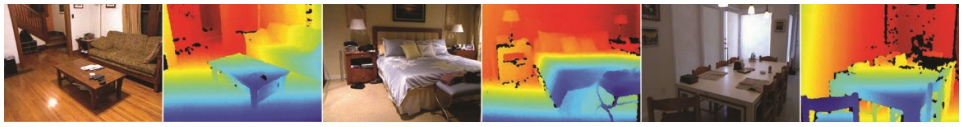


Figure 3. An example of a dataset RGB image and Ground truth depth map. (a) A single RGB image, and (b) the corresponding ground truth.

The coarse network is trained for 2 million samples using stochastic gradient descent (SGD) with batches of size 32. After that, the fixed fine network is trained with 1.5 million samples using outputs from the pre-trained coarse network. The coarse convolutional layers 1-5 have a learning rate of 0.001, while coarse full layers 6 and 7 have a rate of 0.1. Fine layers 1 and 3 are set at 0.001, and fine layer 2 has a rate of 0.01. The ratios were figured out through trial and error on a validation set, which was then included back into the training set for final evaluations, and the overall scale of all the rates was adjusted to a multiple of 5. The momentum is 0.9. The coarse network is trained for 38 hours, and the fine network is trained using a NVidia GTX 1080*Ti* GPU.

4.2. Performance Evaluation

Moreover, to demonstrate the universality of our proposed method, we applied it to the NYU depth v2 dataset and present several examples of depth estimation in **Error! Reference source not found.**. These results illustrate that our method significantly enhances the depth layout of indoor scenes. It is worth emphasizing, the model trained on the NYU v2 dataset can be directly applied to estimate the depth map of similar datasets without requiring fine-tuning.

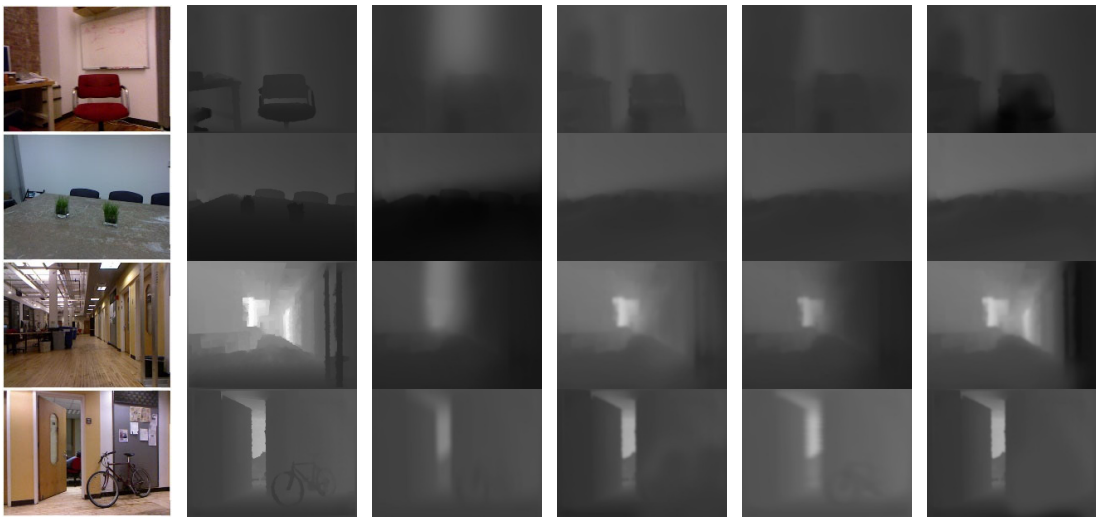


Figure 4. In a visual comparison of the generated depth outcomes, the depth boundaries are anticipated, whereas other methodologies yield hazy and blurry predictions.

Finally, we evaluated the processing speed for estimating the depth map from a single input image and present the results in **Error! Reference source not found.**. For a fair comparison of processing speed, we resized the input image to 512 × 256 pixels and compared our method to two

 doi:10.20944/preprints202411.0868.v1

previous approaches with conditional generative adversarial net Xiaofeng *et al.*, [30], and with Convolutional Neural Networks (CNNs) Gan *et al.*, [31] known for their relatively fast performance. Notably, our proposed method operates exceptionally recklessly. Therefore, we believe that our network architecture has the potential to be widely used in various human-robot interaction system-based applications.

In this outcome, the total number of valid pixels in the ground truth is denoted by T. Based on this, our proposed method compared with state-of-the-art techniques on the NYU v2 dataset [9], and the corresponding outcomes are presented in **Error! Reference source not found.**. Our approach exhibits high performance for all the evaluation metrics. By utilizing latent features from our guided network and corresponding gradients, our proposed method can efficiently restore the depth boundary, thus reducing depth ambiguities and leading to improved performance. In particular, our approach shows a significant improvement in the root mean squared error difference (RMSE) as explained in (4), reducing it by 45.86% compared to the approach proposed by Eigen et al. [32]. Based on these results, it can be concluded that our method can effectively restore the depth layout of indoor scenes, leading to better performance in human-robot interaction systems.

$$RMSE = \sqrt{\frac{1}{|N|}\sum_{y \epsilon T} ||y - y^*||^2} \tag{4}$$

To demonstrate the effectiveness and robustness of the proposed method, a comparison is made with typical methods introduced by Saxena et al. [7]. The study aims to improve the monocular depth estimation by controlling the motion of an RGB camera and capturing images at well-controlled orientations & positions. The results of three methods, Eigen et al. [32], Sihaeng et al.[33], and Zhang et al.[30], are presented in **Error! Reference source not found.** to qualitatively evaluate the performance. Ground truth samples have been incorporated for better visualization.

Table 2. Performance analysis of the state-of-the-art architectures with proposed network architectures.

Architectures	Eigen et al.	Sihaeng et al.	Zhang et al.	Proposed
RMSE	0.907	0.454	0.590	0.416

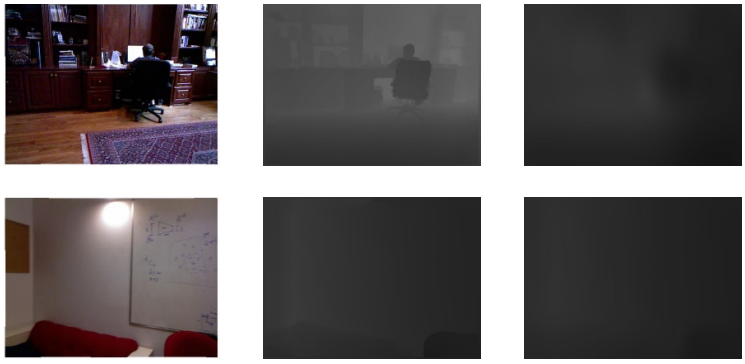
The proposed method successfully restores the depth boundary compared to previous approaches, especially revealing the small sign with high contrast in the first example of **Error! Reference source not found.**, which other methods fail to restore. The boundary of objects is also successfully restored with clear corresponding regions. Based on these comparisons, the proposed method can provide reliable depth information from a single image. The performance is quantitatively evaluated using five metrics: RMSE, which is commonly used for depth estimation performance evaluation.

4.3. Discussion

The issue of depth perception is a crucial aspect of computer vision, which has been the focus of attention of numerous researchers, resulting in significant advances in recent decades. However, most of the work done in this field, such as stereopsis, has relied on using multiple image geometric cues to determine depth. In contrast, single-image cues provide a largely independent source of information, which has not been extensively explored until now. Given the importance of depth and shape perception in various applications, including object recognition, robot grasping, navigation, image compositing, and video retrieval, we believe that monocular depth estimation can significantly enhance these applications, particularly in cases where only a single image of a scene is available. We have developed an algorithm to infer detailed depth estimation from a single still image. Our algorithm surpasses previous methods in both quantitative accuracy and visual quality. Our approach employs a loss function that emphasizes both latent loss and gradient loss. Apart from the assumption that the environment consists of multiple small planes, we do not make any explicit

 doi:10.20944/preprints202411.0868.v1

assumptions about the structure of the scene, unlike Delage et al. [34] and Hoiem et al. [35], who assume that the scene comprises vertical surfaces standing on a horizontal floor. This allows our model to generalize well, even to scenes with significant non-vertical structures.



RGB Input Ground Truth Proposed failed case

In a few situations, our suggested model underperformed and produced results that were completely different from the ground reality. The failure possibilities of the anticipated depth estimate were shown in the figure above. We discovered via our research that the dark, low light, and ground truth without details are difficult to anticipate even closely, thus an environment with defined boundaries is required to process accurate prediction. To demonstrate the effectiveness of our approach in real-world human-robot interaction system applications, the images captured by our proposed technique are not perfectly calibrated and rectified, resulting in more noise in the correspondence of depth estimation. In such cases, single image depth estimation, which is more robust than stereo estimation, is preferred. However, in industrial environments, this approach can be improved for better reconstruction for 3D modeling of indoor environment for human-robot interaction.

5. Conclusions

This paper introduced a new approach for estimating depth from a single monocular image, which can be used for use for reconstructing 3D environments for robots. The main concept of this novel method involves utilizing encoded features extracted from a latent space network to guide the depth estimation process. To improve the quality of the estimated depth map while preserving the sharpness of the depth boundaries, the method utilizes a loss function that emphasizes both latent loss and gradient loss with Residual Blocks which are ampsizing the primary visual perception. It is inspired by V1's role percepting primary visual information of brains to recognize the edges of depth information more clearly. Additionally, the proposed method produces clean boundaries, which makes it suitable for 3D modeling of the scene. Experimental findings demonstrate an outstanding performance of the proposed method in the task of depth estimation to be used for 3D modeling of indoor environment for human-robot interaction.

Author Contributions: The first author participated in (a) conception and design, experimentations, and interpretation of the data; (b) drafting the article or revising it critically, and (c) approval of the final version. The second author supervised this research, conceptualized initial ideas, revised the draft version of the manuscript, and approved the final version. All authors have read and agreed to the published version of the manuscript.

Funding: This Research was supported by the Tongmyong University Research Grants 2022 (2022B001).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Ye, H.; Chen, Y.; Liu, M. Tightly Coupled 3D Lidar Inertial Odometry and Mapping. In Proceedings of the 2019 International Conference on Robotics and Automation (ICRA); IEEE Press: Montreal, QC, Canada, May 20 2019; pp. 3144–3150.

2. Yuan, W.; Hang, K.; Song, H.; Kragic, D.; Wang, M.Y.; Stork, J.A. Reinforcement Learning in Topology-Based Representation for Human Body Movement with Whole Arm Manipulation. In Proceedings of the 2019 International Conference on Robotics and Automation (ICRA); May 2019; pp. 2153–2160.
3. Scharstein, D.; Szeliski, R. A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms. *Int. J. Comput. Vis.* **2002**, *47*, 7–42, doi:10.1023/A:1014573219977.
4. Koenderink, J.J.; Doorn, A.J. van Affine Structure from Motion. *JOSA A* **1991**, *8*, 377–385, doi:10.1364/JOSAA.8.000377.
5. Seitz, S.M.; Curless, B.; Diebel, J.; Scharstein, D.; Szeliski, R. A Comparison and Evaluation of Multi-View Stereo Reconstruction Algorithms. In Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06); June 2006; Vol. 1, pp. 519–528.
6. Stefanoski, N.; Bal, C.; Lang, M.; Wang, O.; Smolic, A. Depth Estimation and Depth Enhancement by Diffusion of Depth Features. In Proceedings of the 2013 IEEE International Conference on Image Processing; September 2013; pp. 1247–1251.
7. Saxena, A.; Sun, M.; Ng, A.Y. Make3D: Learning 3D Scene Structure from a Single Still Image. *IEEE Trans. Pattern Anal. Mach. Intell.* **2009**, *31*, 824–840, doi:10.1109/TPAMI.2008.132.
8. Konrad, J.; Wang, M.; Ishwar, P.; Wu, C.; Mukherjee, D. Learning-Based, Automatic 2D-to-3D Image and Video Conversion. *IEEE Trans. Image Process.* **2013**, *22*, 3485–3496, doi:10.1109/TIP.2013.2270375.
9. Nathan, S.; Derek, H.; Pushmeet, K.; Rob, F. Indoor Segmentation and Support Inference from RGBD Images;
10. Fang, C.-J.; Lin, S.-K. A Performance Criterion for the Depth Estimation of a Robot Visual Control System. In Proceedings of the Proceedings 2001 ICRA. IEEE International Conference on Robotics and Automation (Cat. No.01CH37164); May 2001; Vol. 2, pp. 1201–1206 vol.2.
11. Lopez-Nicolas, G.; Sagues, C.; Guerrero, J.J.; Lopez-Nicolas, G.; Sagues, C.; Guerrero, J.J. *Shortest Path Homography-Based Visual Control for Differential Drive Robots*; IntechOpen, 2007; ISBN 978-3-902613-01-1.
12. Sabnis, A.; Vachhani, L. Single Image Based Depth Estimation for Robotic Applications. In Proceedings of the 2011 IEEE Recent Advances in Intelligent Computational Systems; September 2011; pp. 102–106.
13. Unsupervised Odometry and Depth Learning for Endoscopic Capsule Robots – ArXiv Vanity Available online: https://www.arxiv-vanity.com/papers/1803.01047/ (accessed on 30 March 2023).
14. Jin, Y.; Lee, M. Enhancing Binocular Depth Estimation Based on Proactive Perception and Action Cyclic Learning for an Autonomous Developmental Robot. *IEEE Trans. Syst. Man Cybern. Syst.* **2019**, *49*, 169–180, doi:10.1109/TSMC.2017.2779474.
15. Xiao, B.; Xu, W.; Guo, J.; Lam, H.-K.; Jia, G.; Hong, W.; Ren, H. Depth Estimation of Hard Inclusions in Soft Tissue by Autonomous Robotic Palpation Using Deep Recurrent Neural Network. *IEEE Trans. Autom. Sci. Eng.* **2020**, *17*, 1791–1799, doi:10.1109/TASE.2020.2978881.
16. Cheng, Y.; Yi, P.; Liu, R.; Dong, J.; Zhou, D.; Zhang, Q. Human-Robot Interaction Method Combining Human Pose Estimation and Motion Intention Recognition. In Proceedings of the 2021 IEEE 24th International Conference on Computer Supported Cooperative Work in Design (CSCWD); May 2021; pp. 958–963.
17. Yu, H.; Shen, F. Disparity Estimation Method of Electric Inspection Robot Based on Lightweight Neural Network. In Proceedings of the 2021 6th International Conference on Intelligent Computing and Signal Processing (ICSP); April 2021; pp. 929–932.
18. Wang, L.; Zhang, J.; Wang, Y.; Lu, H.; Ruan, X. CLIFFNet for Monocular Depth Estimation with Hierarchical Embedding Loss. In Proceedings of the Computer Vision – ECCV 2020; Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M., Eds.; Springer International Publishing: Cham, 2020; pp. 316–331.
19. Shimada, N.; Shirai, Y.; Kuno, Y.; Miura, J. Hand Gesture Estimation and Model Refinement Using Monocular Camera-Ambiguity Limitation by Inequality Constraints. In Proceedings of the Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition; April 1998; pp. 268–273.
20. Pan, H.; Huang, J.; Qin, S. High Accurate Estimation of Relative Pose of Cooperative Space Targets Based on Measurement of Monocular Vision Imaging. *Optik* **2014**, *125*, 3127–3133, doi:10.1016/j.ijleo.2013.12.020.
21. Learning Latent Vector Spaces for Product Search | Proceedings of the 25th ACM International on Conference on Information and Knowledge Management Available online: https://dl.acm.org/doi/10.1145/2983323.2983702 (accessed on 30 March 2023).
22. Kashyap, H.J.; Fowlkes, C.; Krichmar, J.L. Sparse Representations for Object and Ego-Motion Estimation in Dynamic Scenes. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, *32*, 2521–2534, doi:10.1109/TNNLS.2020.3006467.
23. Reading, C.; Harakeh, A.; Chae, J.; Waslander, S.L. Categorical Depth Distribution Network for Monocular 3D Object Detection.; IEEE Computer Society, June 1 2021; pp. 8551–8560.
24. Pei, M. MSFNet:Multi-Scale Features Network for Monocular Depth Estimation 2021.
25. Zhao, C.; Tang, Y.; Sun, Q. Unsupervised Monocular Depth Estimation in Highly Complex Environments 2022.

26. Guo, S.; Rigall, E.; Ju, Y.; Dong, J. 3D Hand Pose Estimation From Monocular RGB With Feature Interaction Module. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *32*, 5293–5306, doi:10.1109/TCSVT.2022.3142787.
27. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015; Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F., Eds.; Springer International Publishing: Cham, 2015; pp. 234–241.
28. Johnson, J.; Alahi, A.; Fei-Fei, L. Perceptual Losses for Real-Time Style Transfer and Super-Resolution. In Proceedings of the Computer Vision – ECCV 2016; Leibe, B., Matas, J., Sebe, N., Welling, M., Eds.; Springer International Publishing: Cham, 2016; pp. 694–711.
29. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); June 2016; pp. 770–778.
30. Zhang, X.; Chen, S.; Xu, Q.; Zhang, X. Monocular Image Depth Estimation Using a Conditional Generative Adversarial Net. In Proceedings of the 2018 37th Chinese Control Conference (CCC); July 2018; pp. 9176–9180.
31. Gan, Y.; Xu, X.; Sun, W.; Lin, L. Monocular Depth Estimation with Affinity, Vertical Pooling, and Label Enhancement. In Proceedings of the Computer Vision – ECCV 2018; Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y., Eds.; Springer International Publishing: Cham, 2018; pp. 232–247.
32. Depth Map Prediction from a Single Image Using a Multi-Scale Deep Network | Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2 Available online: https://dl.acm.org/doi/10.5555/2969033.2969091 (accessed on 30 March 2023).
33. Patch-Wise Attention Network for Monocular Depth Estimation(AAAI 2020) - KAIST Available online: https://ee.kaist.ac.kr/en/ai-in-signal/18520/ (accessed on 30 March 2023).
34. Delage, E.; Lee, H.; Ng, A.Y. A Dynamic Bayesian Network Model for Autonomous 3D Reconstruction from a Single Indoor Image. In Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06); June 2006; Vol. 2, pp. 2418–2428.
35. Hoiem, D.; Efros, A.A.; Hebert, M. Geometric Context from a Single Image. In Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1; October 2005; Vol. 1, pp. 654-661 Vol. 1.