

A deep learning based stereo matching model for autonomous vehicle

Deepa¹, Jyothi Kupparu²

¹Department of Information Science and Engineering, Nitte Mahalinga Adyanthaya Memorial Institute of Technology-Affiliated to Nitte (Deemed to be University), Nitte, India

²Department of Information Science and Engineering, Jawaharlal Nehru National College of Engineering, Visvesvaraya Technological University, Shimoga, India

Article Info

Article history:

Received Dec 23, 2021

Revised Jul 23, 2022

Accepted Aug 21, 2022

Keywords:

Convolutional neural networks

Disparity

Generative adversarial network

Ill posed regions

Stereo matching

ABSTRACT

Autonomous vehicle is one the prominent area of research in computer vision. In today's AI world, the concept of autonomous vehicles has become popular largely to avoid accidents due to negligence of driver. Perceiving the depth of the surrounding region accurately is a challenging task in autonomous vehicles. Sensors like light detection and ranging can be used for depth estimation but these sensors are expensive. Hence stereo matching is an alternate solution to estimate the depth. The main difficulties observed in stereo matching is to minimize mismatches in the ill-posed regions, like occluded, texture less and discontinuous regions. This paper presents an efficient deep stereo matching technique for estimating disparity map from stereo images in ill-posed regions. The images from Middlebury stereo data set are used to assess the efficacy of the model proposed. The experimental outcome depicts that the proposed model generates reliable results in the occluded, texture less and discontinuous regions as compared to the existing techniques.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Deepa

Department of Information Science and Engineering, Nitte Mahalinga Adyanthaya Memorial Institute of Technology-Affiliated to Nitte (Deemed to be University)

Nitte, India

Email: deepashetty17@nitte.edu.in

1. INTRODUCTION

Autonomous vehicles are a prominent research topic in the computer vision. It is necessary to correctly measure the three-dimensional (3D) view of the surrounding region of the vehicle in real time to make a driving decision. Precision of the depth map is crucial for the safety measure of autonomous vehicles. In these vehicles the depth details of the surrounding region are usually extracted using the hardware like light detection and ranging sensors. These sensors are expensive to install and also have certain drawbacks that may lower the standard of the depth information. These sensors do not provide additional information like traffic light color which plays a major role in decision making. Computer vision based stereo matching could be an alternate solution to overcome this drawback. The aim of stereo matching is to find matching pixels of images from different viewpoints and then estimate the depth [1]–[3]. It finds its applications in augmented reality, robotics, 3D reconstruction [4]–[9]. Stereo vision tries to imitate the process in the human eye and the human brain. A scene taken from two cameras displaced horizontally will form two slightly separate projections. Disparity is the horizontal displacement in an object. A map that contains displacement of all pixels in an image is known as disparity map. Depth of a scene can be estimated from this disparity map.

In recent past many stereo algorithms were proposed [10]–[12]. A classic stereo algorithm mainly follows three steps namely: computing the pixel wise features, construction of cost volume followed by post-processing. Traditional stereo matching methods are grouped as local, global and semi-global methods. Local methods rely on low level pixel features to compute the similarity in the cost computation step. They estimate the correspondence by means of a window or support region [13]–[15]. Since the pixel wise characterization play a major factor, a wide variety of these representations are used by researchers varying from a simple rgb representation of pixels to the other descriptors like census transform, scale invariant feature transform. Segment based super pixel technique is proposed in [16]. After finding the edges and matching cost, adaptive support weight is used in cost aggregation. It proposes dual path refinement to correct disparities. Stereo matching based on adaptive cross area and guided filtering with orthogonal weights (ACR-GIF-OW) is proposed in [17]. These techniques are computationally less expensive but do not produce accurate results in the texture less, discontinuous and occluded areas.

A Global methods handle texture less regions or uneven surfaces by including smoothness cost. Global methods make use of global energy function. The energy function is minimized step by step to compute disparity by assuming matching as a labelling problem. The pixels are considered as nodes and disparity estimated is considered as labels. The global methods use data and smoothness term to compute the energy function to produce smooth disparity. Graph cut [18], dynamic programming [19] and belief propagation [20] are the classic global matching algorithm. A tree structure is proposed in [21] named pyramid-tree that performs cross regional smoothing and handling region of low texture. In addition, they used log angle for cost computation which is robust to inconsistencies. The performance of global methods is limited because these approaches depend on hand-crafted features and hence do not produce accurate results.

Convolutional neural network (CNN) is popular in different vision [22]–[24] applications. These methods are widely used in stereo matching. It improves the performance as compared to traditional methods. Kendall *et. al.* [25] the authors presented an architecture that learns disparity without regularization. Features are extracted automatically using CNN without any manual intervention. These features are used to perform stereo matching, that can handle texture less regions or uneven surfaces. Eigen *et. al.* [26] made use of basic neural networks to determine depth of a scene. They used AlexNet architecture to generate coarse map. Another network is followed that performs local refinements. The work proposed in [27] included the process of multi-stage framework that combined random forests and CNN. An architecture named neural regression forest is used to find depth from single input image. It allows parallel training of all CNN. Finally, a bilateral filter was used to obtain a refined disparity map. A similar concept is presented in [28] where many tiny neural networks were trained across overlapping patches. DispNet is one of the basic networks used for disparity estimation. A cascading residual learning network is used in [29] that extend the DispNet structure. It is obtained by using DispFullNet and DispResNet. The initial stages of CNN uses DispNet with an additional up convolution module. This help to extract more information. The next stage generates residual signal that helps in refinement. A trainable network is explained in [30]. It uses a robust differentiable patch match internal structure that discards most disparities without performing cost volume evaluation fully. This reduces search space and increases memory and time efficiency. The main drawbacks of existing methods are that the ill posed regions are not handled effectively. In the proposed method CNN is combined with optimization technique. CNN is used to replace the the hand-crafted term with the learned features. The output of CNN is used to calculate the unary and smoothness cost. Smoothness cost is added by taking the information from the neighboring pixels. Smoothness cost estimates the contrast-sensitive information to get a smooth disparity map. Post processing is performed to handle occlusion.

In stereo vision, the areas visible in one view may not be visible in another. It is often difficult to reconstruct such regions in one image by looking at the other. The losses computed in these areas are noisy, leading to inaccurate results specifically in the occluded areas. Disparity refinement is implemented to enhance the accuracy of matching in ill posed areas. The left-right consistency check is the common method used to identify and handle the outliers. Even though several methods were proposed in the past to enhance the efficiency of matching, the low accuracy problem especially in the ill posed areas has not been handled very well. In order to handle these areas, post processing is performed by means of a generative adversarial network (GAN) model put forward by Goodfellow [31]. GAN is a structure used for training generative model. It uses the concept of min-max game. The two models namely generator model and a discriminative model is used to analyze the distribution of data. The generator tries to understand the distribution which is almost same to the real distribution of data. The ability to generate high quality image by GAN makes it applicable in several image processing applications. An encoder decoder structure is used for training in reconstructing the images. This model can produce various realistic representation of input by altering the attribute values. A conditional adversarial network [32] can be used for image translation. This translation converts the image from one representation to the other such as day to night.

We propose a hybrid CNN based deep stereo network model (CDSN) to estimate the disparity map that can produce accurate results. Loopy belief propagation is used to compute initial disparity map from features extracted from CNN. A generative neural network is used to handle the ill posed regions in the disparity map. The generated images look more realistic and closer to ground truth disparity map. The obtained result show that the proposed CDSN model handle the ill posed regions like discontinuities in the image boundaries and occluded areas effectively. The proposed model outperforms the other existing techniques on Middlebury dataset [33]. The paper is organized in a manner, section 2 explains the proposed CDSN model. Section 3 depicts the results of proposed model. The conclusions of the paper are presented in section 4.

2. METHOD

A CNN based model is proposed for stereo matching to find disparity map. The features extracted from CNN is used to compute the unary cost and smoothness cost. Global energy function is adapted to get the initial disparity map. A GAN model is used to handle ill posed region. Table 1 depicts the list of symbols with its description. The flow chart of proposed model is displayed in Figure 1.

Table 1. Symbol table

Symbol	Description
$D(d_i)$	Unary cost at pixel 'i'
V_i^l	Left feature vector
V_i^r	Right feature vector
$S(d_i, d_j)$	Smoothness cost
α and β	Smoothness constants
$msg_{i \rightarrow j}(d_i)$	Message from pixel 'i' to 'j' at iterations 't'
$D(p, q)$	discriminator
$E_{p,q}$	Expected values of all real data instances
$G(p,r)$	generator
$E_{p,r}$	Expected values of all generated instances
D_g	Ground truth disparity map
D_t	Estimated disparity map
T	Threshold value
N	Total number of pixels

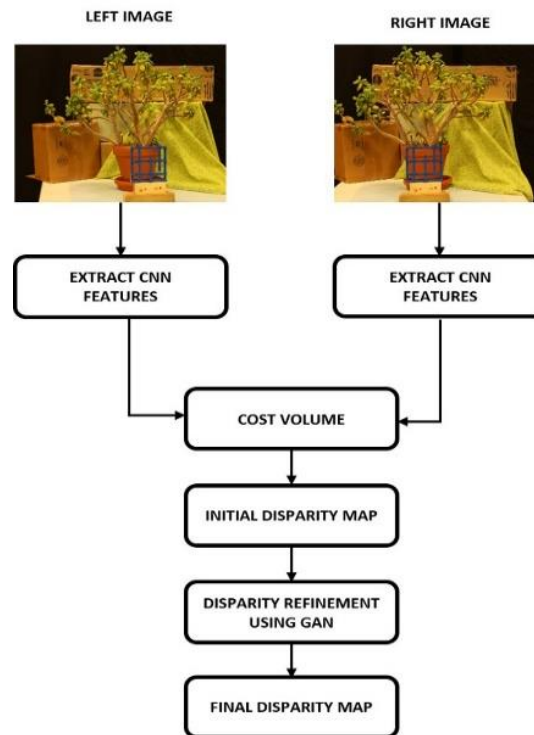


Figure 1. Flowchart of the CDSN model

2.1. CNN feature extraction

Conventional algorithms for stereo matching focuses on hand crafted features which leads to inadequate image information. CNN is used for the various vision problems including stereo matching. The CNN can extract local context better, hence it is robust to any photometric differences. The feature descriptors are extracted from rectified stereo images using a pre-trained visual geometry group (VGG-16) model [34]. The VGG-16 model is trained using ImageNet dataset which contains 14 million labelled images that are of high resolution that belong to 1,000 classes. The output of the 9th layer is used for stereo matching in the proposed model as it presents an appropriate feature space for computing disparity. VGG-16 uses a max pool layer that select the maximum element from the input map using a filter of 2×2 . The first and second layers include 64 channels of 3×3 kernel size which is followed by max pool function of stride 2, 2. The third and fourth layers include 128 channels of 3×3 kernel followed by max pool function of stride 2, 2. The next three layers include 256 channels of 3×3 kernel that is followed by max pool function of stride 2, 2. Eighth and ninth layers include 512 channels of 3×3 kernel size. An N-dimensional feature vector is obtained for every location of pixel.

2.2. Initial disparity map estimation

The extracted feature descriptors are used to determine the matching cost of every pixel in left feature map. We search horizontally along the right feature map for the best matching value. The matching unary cost is calculated using the Euclidian distance of two feature descriptors using (1).

$$D(d_i) = \min \|V^l - V^r\| \quad (1)$$

Unary cost may not yield optimal result in the texture less, repetitive patterns, discontinuity regions. The smoothness cost is used to smoothen the unary cost. Many smoothening techniques is proposed in the recent past. Most of these methods use random variables to have the disparity of a pixel, which encodes smoothness cost based on some standard constant. The smoothness cost is estimated based on neighbouring pixel information. The smoothness cost penalizes the inconsistent disparity values. The smoothness cost is computed using (2),

$$S(d_i, d_j) = \frac{\alpha * (d_i - d_j)^2}{(d_i - d_j)^2 + \beta} \quad (2)$$

Let P represent pixels in the image. The initial disparity map d_i of each pixel $i \in P$ is estimated using energy function E

$$E(d) = \sum_{i \in P} D(d_i) + \sum_{(i,j) \in N} S(d_i, d_j) \quad (3)$$

The proposed method uses max product variation of loopy belief propagation (LBP) [20] to obtain the best disparity map. LBP is an algorithm based on assigning label to each pixel imposing global constraints and message passing. This is an iterative method where the messages are passed to left, right, top and bottom in each iteration. In each iteration t , the message is passing from pixel i to pixel j using (4),

$$msg_{i \rightarrow j}^t(d_i) = \min_{d_i} [D(d_i) + S(d_i, d_j) + \sum_{a \in N(i) \setminus j} msg_{a \rightarrow i}^{t-1}(d_i)] \quad (4)$$

Here a represents all neighbours of i except j
Belief is calculated by (5).

$$Belief(d_i) = D(d_i) + \sum_{k \in N(i)} msg_{k \rightarrow i}^T(d_i) \quad (5)$$

The values d ranges from 0 to maximum disparity range and k represent neighbours of pixel i . The smooth disparity is obtained for iteration T that minimizes the $Belief(d_i)$. It is observed that the minimization of energy became constant after 10 iterations. Hence the proposed algorithm used 10 iterations.

2.3. Disparity refinement using GAN

The GAN network is used to refine the disparity. This refinement model is used to handle ill posed regions. The GAN can perform learning task automatically by identifying various patterns or irregularities from the input data. GANs have the ability to handle missing data such as occluded pixels in the disparity map. The two sub models in GAN are generator and discriminator. The generator model generates new

samples and discriminator model checks if the generated samples are similar to ground truth map. In the proposed model the network learns through ground truth. The architecture of this refinement technique is given in Figure 2.

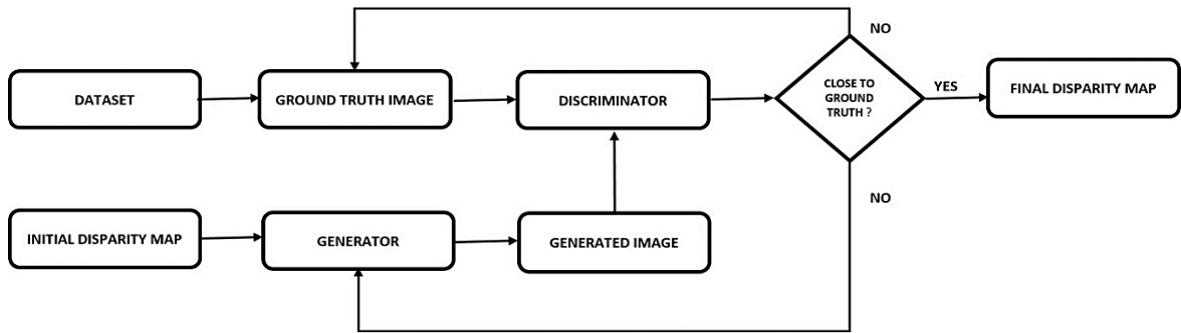


Figure 2. Architecture of disparity refinement technique

The proposed model uses Pix2Pix GAN model [32]. Pix2Pix GAN is simple and can produce high quality images for image translation applications. The efficiency of this GAN as compared to other GAN like CycleGAN [35] and DualGAN [36] is explained in the ablation study. The generator in Pix2Pix is a convolutional network that accepts initial disparity map as the input image and passes it through several convolution and up-sampling layers. Finally, it produces a refined disparity map, where all the occluded areas are filled with valid data. The U-Net auto encoding generator model is trained using adversarial loss that encourages it to create reasonable image. The encoder and decoder are made up of blocks of convolutional, activation layers and batch normalization layers. The generator is updated by loss that is generated between generated image and ground truth image. This information helps generator model to create more reasonable image that is similar to ground truth. The generator G is trained so as to generate output which can be differentiated from ground truth image by a discriminator D . The GAN objective is represented as,

$$L_{GAN}(G, D) = E_{p,q} [\log D(p, q)] + E_{p,r} [\log (1 - D(p, G(p, r)))] \quad (6)$$

Here p denote a ground truth image, q represent the generated image and r represent the initial disparity map

$$G^* = \operatorname{argmin}_g \max_d L_{GAN}(G, D) \quad (7)$$

G aims to decrease the objective and D aims to increase the objective.

The generator G tries to move the generated image closer to ground truth image using loss L_1 which is calculated as

$$Loss_{L_1}(G) = E_{p,q,r} [\|q - G(p, r)\|_1] \quad (8)$$

The final objective is represented as

$$G^* = \operatorname{argmin}_g \max_d L_{GAN}(G, D) + \lambda Loss_{L_1}(G) \quad (9)$$

The visual arti-facts were reduced for the value of $\lambda=100$.

The network is trained by images from the Middlebury dataset [33]. The network is tested for 100, 200, 300, 400 epochs. The best disparity map is achieved for 300 epochs. The output from the generator is fed to the discriminator together with ground truth image. The gradient loss is calculated with respect to generator and discriminator to update the model. The trained model is tested to yield a best disparity map. It is observed from the results that best disparity map was obtained by handling the ill posed regions. Figure 3 shows the performance of the model with respect to training loss and training accuracy. Figure 3(a) depicts training loss and training accuracy against the number of epochs is shown in Figure 3(b). Lower the loss better is the accuracy.

To measure the efficacy of the model proposed, we deployed and tested our model on Dual Intel Xeon E5-2609V4 8C 1.7 GHz 20M 6.4 GT/s with 128GB memory, Dual NVIDIA Tesla P100 graphics

processing unit (GPU) with 3584 cores and maximum of 18.7 TeraFLOPS. The proposed CDSN model is evaluated on Middlebury dataset images. These images are pre-processed and rectified stereo images. The output of the 9th layer pre-trained VGG-16 architecture is used for estimating initial disparity map using loopy belief propagation. Initial disparity map is estimated using python programming. GAN is implemented using Pytorch. The Adam optimizer is used to train the Pix2Pix GAN for 300 epochs to handle the ill posed regions. The learning rate has been initialised to 0.0002. The complexity of GAN model is summarized in the Table 2.

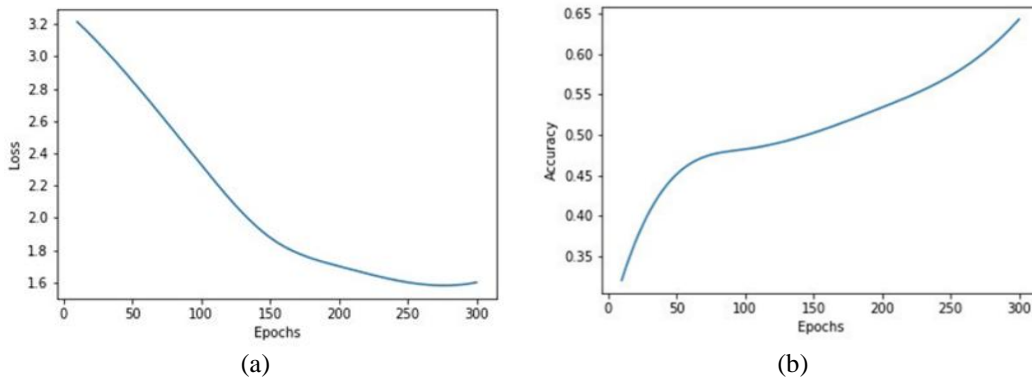


Figure 3. Performance of the model (a) training loss versus number of epochs and (b) training accuracy versus number of epochs

Table 2. Complexity of GAN model

Input size	Optimizer	Parameters	Epochs	Output size	GPU memory	GPU model
256X256X3	Adam	54.414M	300	256X256X3	128GB	Dual NVIDIA Tesla P100

3. RESULTS AND DISCUSSION

The proposed model is analyzed for the images taken from Middlebury datasets namely “Jade plant”, “Piano”, “Pipes”, and “Recycle”. The test images with resolution are shown in Table 3. Middlebury 2014 dataset contains 33 scenes that are classified into training, additional images and test images. Certain images are used more than once under various exposure. A very high-resolution images is the salient feature of the dataset. Ground truth maps and images are given at quarter, half and full resolution.

Table 3. Images from Middlebury 2014

Images	Image resolution
Jade plant	659x497
Piano	707x481
Pipes	735X485
Recycle	720x486

3.1. Qualitative comparison

The qualitative results for estimating disparity map is depicted in Figure 4. From the top to bottom: Jade plant, piano, pipes, and recycle. Figure 4(a) shows the left image, Figure 4(b) shows the right image, Figure 4(c) represent the ground truth image and Figure 4(d) represent the estimated disparity map.

3.2. Quantitative comparison

The percentage of bad matching pixel (PBMP) and root mean square error (RMSE) metrics were used for quantitative analysis. Lower values of PBMP and RMSE indicates better efficiency. PBMP is calculated,

$$PBMP = \left[\frac{1}{N} \sum |d_t(x, y) - d_g(x, y)| > T \right] * 100 \quad (10)$$

RMSE is calculated as,

$$RMSE = \left[\frac{1}{N} \sum |d_t(x, y) - d_g(x, y)|^2 \right]^{\frac{1}{2}} \quad (11)$$

For evaluations purpose, we compared CDSN model with existing stereo matching model. The compared matching models are: deep pruner [30] ACR-GIF-OW [17], and efficient stereo matching by log-angle and pyramid-tree (LPSM) [21]. The occluded areas are not dealt efficiently in [30]. Stereo matching proposed in [17] is computationally less expensive but do not produce accurate results in the texture less, discontinuous areas. Stereo matching proposed in [21] rely on hand-crafted cost matching and hence results produced are not accurate. The Middlebury evaluation leader board results of existing methods are used for comparison. The PBMP and RMSE results of the proposed model and existing techniques are shown in Table 4 and Table 5 respectively. The PBMP and average RMSE results of the proposed CDSN model is less than all three compared method. Hence the proposed model outperforms the compared method and hence suitable for disparity map estimation.

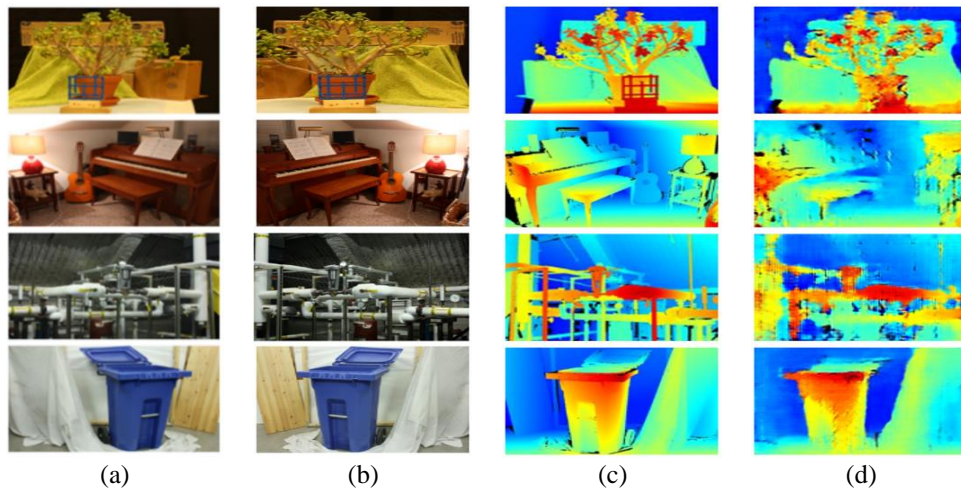


Figure 4. Visual results on Middlebury images (a) left image, (b) right image, (c) ground truth image and (d) estimated disparity map

Table 4. The quantitative results based on PBMP for error threshold = 1 between computed and ground truth disparities

	Jade plant	Piano	Pipes	Recycle
DEEP PRUNER [30]	62.8	41.0	53.8	36.8
ACR-GIF-OW [17]	51.8	45.1	40.5	37.5
LPSM [21]	59.2	44.8	46.3	36.8
CDSN	50.76	39.71	40.47	33.57

Table 5. The quantitative results based on RMSE between computed and ground truth disparities

	Jade plant	Piano	Pipes	Recycle	Average
DEEP PRUNER [25]	28.2	4.64	13.7	3.81	12.58
ACR-GIF-OW [17]	64.9	14.8	28.6	15.8	31.02
LPSM [21]	34.8	6.09	16.3	5.79	15.74
CDSN	6.07	8.73	8.88	8.48	8.04

3.3. Ablation study

We executed ablation study by comparing the proposed model with the models like CycleGAN and DualGAN. CycleGAN is a technique that performs image translation without using paired examples. This GAN uses unsupervised training. DualGAN is made up of two generators and two discriminators. It is trained to translate images from source to target and target to source. The various metric used are absolute relative distance (ARD), squared relative difference (SRD) and RMSE. Lower values indicate better performance. We find the efficiency of the proposed model is significantly high which is presented in the Table 6.

$$ARD = \frac{1}{N} \sum \frac{d_t(x,y) - d_g(x,y)}{d_t(x,y)} \quad (12)$$

$$SRD = \frac{1}{N} \sum \frac{|d_t(x,y) - d_g(x,y)|^2}{d_t(x,y)} \quad (13)$$

Table 6. Ablation study using metrics ARD, SRD, RMSE

	CycleGAN	DualGAN	Proposed model	
ARD	0.032	0.035	0.016	Lower is better
SRD	0.352	0.374	0.337	
RMSE	7.036	7.974	6.591	

4. CONCLUSION

This paper presents a novel CNN based model for stereo matching to estimate disparity map from rectified stereo images which is useful in autonomous vehicles. The features extracted from CNN is used to compute the unary cost and smoothness cost. The initial disparity map is obtained using loopy belief propagation, which is then refined using a GAN model to handle the ill posed regions. It is found that the proposed model based on CNN generated disparity maps which are smoother than those generated using naive model and the ill posed regions are handled well using GAN network. The proposed model is evaluated qualitatively as well as quantitatively on various images from Middlebury stereo data set. The results determine that proposed model achieves best disparity map and outperforms existing methods.




REFERENCES

- [1] D. Scharstein, R. Szeliski, and R. Zabih, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *Int. J. Comput. Vis.*, no. 1, pp. 7–42, 2002, doi: 10.1109/SMBV.2001.988771.
- [2] R. A. Hamzah and H. Ibrahim, "Literature survey on stereo vision disparity map algorithms," *Journal of Sensors*, vol. 2016, 2016, doi: 10.1155/2016/8742920.
- [3] M. S. Hamid, N. F. A. Manap, R. A. Hamzah, and A. F. Kadmin, "Stereo matching algorithm based on deep learning: A survey," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 5, pp. 1663–1673, 2022, doi: 10.1016/j.jksuci.2020.08.011.
- [4] U. Hani and L. Moin, "Realtime autonomous navigation in V-Rep based static and dynamic environment using EKF-SLAM," *IAES International Journal of Robotics and Automation (IJRA)*, vol. 10, no. 4, p. 296, 2021, doi: 10.11591/ijra.v10i4.pp296-307.
- [5] Susanto, D. D. Budiarjo, A. Hendrawan, and P. T. Pungkasanti, "The implementation of intelligent systems in automating vehicle detection on the road," *IAES International Journal of Artificial Intelligence*, vol. 10, no. 3, pp. 571–575, 2021, doi: 10.11591/ijai.v10.i3.pp571-575.
- [6] S. Sivaraman and M. M. Trivedi, "A review of recent developments in vision-based vehicle detection," *IEEE Intelligent Vehicles Symposium, Proceedings*, pp. 310–315, 2013, doi: 10.1109/IVS.2013.6629487.
- [7] S. Hong, M. Li, M. Liao, and P. Van Beek, "Real-time mobile robot navigation based on stereo vision and low-cost GPS," *IS and T International Symposium on Electronic Imaging Science and Technology*, pp. 10–15, 2017, doi: 10.2352/ISSN.2470-1173.2017.9.IRIACV-259.
- [8] B. Krajancich, P. Kellnhofer, and G. Wetzstein, "Optimizing depth perception in virtual and augmented reality through gaze-contingent stereo rendering," *ACM Transactions on Graphics*, vol. 39, no. 6, 2020, doi: 10.1145/3414685.3417820.
- [9] H. Ham, J. Wesley, and Hendra, "Computer vision based 3D reconstruction : a review," *International Journal of Electrical and Computer Engineering*, vol. 9, no. 4, pp. 2394–2402, 2019, doi: 10.11591/ijece.v9i4.pp2394-2402.
- [10] C. Bai, Q. Ma, P. Hao, Z. Liu, and J. Zhang, "Improving stereo matching algorithm with adaptive cross-scale cost aggregation," *International Journal of Advanced Robotic Systems*, vol. 15, no. 1, 2018, doi: 10.1177/1729881417751544.
- [11] H. Shabanian and M. Balasubramanian, "A new hybrid stereo disparity estimation algorithm with guided image filtering-based cost aggregation," *IS and T International Symposium on Electronic Imaging Science and Technology*, vol. 2021, no. 2, 2021, doi: 10.2352/ISSN.2470-1173.2021.2.SDA-059.
- [12] R. A. Hamzah, M. G. Y. Wei, and N. S. N. Anwar, "Development of stereo matching algorithm based on sum of absolute RGB color differences and gradient matching," *International Journal of Electrical and Computer Engineering*, vol. 10, no. 3, pp. 2375–2382, 2020, doi: 10.11591/ijece.v10i3.pp2375-2382.
- [13] H. Liu, R. Wang, Y. Xia, and X. Zhang, "Improved cost computation and adaptive shape guided filter for local stereo matching of low texture stereo images," *Applied Sciences (Switzerland)*, vol. 10, no. 5, 2020, doi: 10.3390/app10051869.
- [14] S. Chen, J. Zhang, and M. Jin, "A simplified ICA-based local similarity stereo matching," *Visual Computer*, vol. 37, no. 2, pp. 411–419, 2021, doi: 10.1007/s00371-020-01811-x.
- [15] J. K. and P. C. J., "Multi modal face recognition using block based curvelet features," *International Journal of Computer Graphics & Animation*, vol. 4, no. 2, pp. 21–37, 2014, doi: 10.5121/ijcga.2014.4203.
- [16] C. S. Huang, Y. H. Huang, D. Y. Chan, and J. F. Yang, "Shape-reserved stereo matching with segment-based cost aggregation and dual-path refinement," *Eurasip Journal on Image and Video Processing*, vol. 2020, no. 1, 2020, doi: 10.1186/s13640-020-00525-3.
- [17] L. Kong, J. Zhu, and S. Ying, "Local stereo matching using adaptive cross-region-based guided image filtering with orthogonal weights," *Mathematical Problems in Engineering*, vol. 2021, 2021, doi: 10.1155/2021/5556990.
- [18] V. Kolmogorov, P. Monasse, and P. Tan, "Kolmogorov and zabih's graph cuts stereo matching algorithm," *Image Processing On Line*, vol. 4, pp. 220–251, 2014, doi: 10.5201/ipol.2014.97.
- [19] O. Veksler, "Stereo correspondence by dynamic programming on a tree," *Proceedings - 2005 IEEE Computer Society Conference*




- on *Computer Vision and Pattern Recognition, CVPR 2005*, vol. II, pp. 384–390, 2005, doi: 10.1109/CVPR.2005.334.
- [20] J. Sun, H. Y. Shum, and N. N. Zheng, “Stereo matching using belief propagation,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 2351, pp. 510–524, 2002, doi: 10.1007/3-540-47967-8_34.
- [21] C. Xu, C. Wu, D. Qu, F. Xu, H. Sun, and J. Song, “Accurate and efficient stereo matching by log-angle and pyramid-tree,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 10, pp. 4007–4019, 2021, doi: 10.1109/TCSVT.2020.3044891.
- [22] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017, doi: 10.1145/3065386.
- [23] M. S. Hamid, N. A. Manap, R. A. Hamzah, A. F. Kadmin, S. F. A. Gani, and A. I. Herman, “A new function of stereo matching algorithm based on hybrid convolutional neural network,” *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 25, no. 1, pp. 223–231, 2022, doi: 10.11591/ijeecs.v25.i1.pp223-231.
- [24] E. Shelhamer, J. Long, and T. Darrell, “Fully convolutional networks for semantic segmentation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 640–651, 2017, doi: 10.1109/TPAMI.2016.2572683.
- [25] A. Kendall *et al.*, “End-to-end learning of geometry and context for deep stereo regression,” *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2017-Octob, pp. 66–75, 2017, doi: 10.1109/ICCV.2017.17.
- [26] 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. 3, no. January, pp. 2366–2374, 2014.
- [27] A. Roy and S. Todorovic, “Monocular depth estimation using neural regression forest,” *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2016-Decem, pp. 5506–5514, 2016, doi: 10.1109/CVPR.2016.594.
- [28] A. Chakrabarti, J. Shao, and G. Shakhnarovich, “Depth from a single image by harmonizing overcomplete local network predictions,” *Advances in Neural Information Processing Systems*, pp. 2666–2674, 2016.
- [29] J. Pang, W. Sun, J. S. J. Ren, C. Yang, and Q. Yan, “Cascade residual learning: a two-stage convolutional neural network for stereo matching,” *Proceedings - 2017 IEEE International Conference on Computer Vision Workshops, ICCVW 2017*, vol. 2018-Janua, pp. 878–886, 2017, doi: 10.1109/ICCVW.2017.108.
- [30] S. Duggal, S. Wang, W. C. Ma, R. Hu, and R. Urtasun, “Deeppruner: learning efficient stereo matching via differentiable patchmatch,” *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2019-Octob, pp. 4383–4392, 2019, doi: 10.1109/ICCV.2019.00448.
- [31] I. J. Goodfellow *et al.*, “Generative adversarial nets,” *Advances in Neural Information Processing Systems*, vol. 3, no. January, pp. 2672–2680, 2014, doi: 10.3156/jsoft.29.5_177_2.
- [32] P. Isola, J. Y. Zhu, T. Zhou, and A. A. Efros, “Image-to-image translation with conditional adversarial networks,” *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, vol. 2017-Janua, pp. 5967–5976, 2017, doi: 10.1109/CVPR.2017.632.
- [33] D. Scharstein *et al.*, “High-resolution stereo datasets with subpixel-accurate ground truth,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 8753, pp. 31–42, 2014, doi: 10.1007/978-3-319-11752-2_3.
- [34] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 2015.
- [35] J. Y. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired image-to-image translation using cycle-consistent adversarial networks,” *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2017-Octob, pp. 2242–2251, 2017, doi: 10.1109/ICCV.2017.244.
- [36] Z. Yi, H. Zhang, P. Tan, and M. Gong, “DualGAN: unsupervised dual learning for image-to-image translation,” *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2017-Octob, pp. 2868–2876, 2017, doi: 10.1109/ICCV.2017.310.

BIOGRAPHIES OF AUTHORS



Deepa    is a graduate with M. Tech in Computer Science Engineering from N.M.A.M. Institute of Technology, Nitte, India. She is currently pursuing her Ph.D. degree in Computer Science engineering at VTU. She is currently working as an assistant professor at Information Science & Engineering in N.M.A.M. Institute of Technology, Nitte, India. Her research interests are in fields of computer vision, digital image processing. She has published several papers in international journals and conferences. She can be contacted at email: deepashetty17@nitte.edu.in.



Jyothi Kupparu    received the Ph. D degree in computer science from the Kuvempu University, Shimoga, India. She is a Professor of Information Science & Engineering at J.N.N.C.E Shimoga. Her research interests include image processing, stereo correspondence algorithms for face images, multimodal face recognition, 3D sparse reconstruction and techniques based on stereo rectification for face images. She has published several papers in international journals and conferences. She can be contacted at email: jyothik@jnnce.ac.in.