An Improved Depth Estimation using Stereo Matching and Disparity Refinement Based on Deep Learning

Deepa[1], Jyothi K[2], Abhishek A Udupa[3]
N.M.A.M Institute of Technology, Affiliated to NITTE (Deemed to be University), Nitte, Karkala,
Visvesveraya Technological University, Belagavi, Karnataka, India[1]
J.N.N College of Engineering, Shimoga, Visvesveraya Technological University, Belagavi, Karnataka, India[2]

***Abstract*—Stereo matching techniques are a vital subject in computer vision. It focuses on finding accurate disparity maps that find its use in several applications namely reconstruction of a 3D scene, navigation of robot, augmented reality. It is a method of obtaining corresponding matching point in stereo images to get disparity map. With additional details, this disparity map could be converted into a depth of a scene. Obtaining an efficient disparity map in the texture less, occluded, and discontinuous areas is a difficult job. A matching cost using an improvised Census transform and an optimization framework is proposed to produce an initial disparity map. The classic Census transform focus on the value of pixel at the center. If this pixel is prone to noisy condition, then the census encoding may differ which leads to mismatches. To overcome this issue an improved census transform based on weighted sum values of the neighborhood pixels is proposed which suppresses the noise during stereo matching. Additionally, a deep learning based disparity refinement technique using the generative adversarial network to handle texture less, occluded, and discontinuous areas is proposed. The suggested method offers cutting-edge performance in terms of both qualitative and quantitative outcomes.**

Keywords*—Census transform; deep learning; depth; generative adversarial network; occlusion; stereo matching

I. Introduction

Stereo matching has gathered attraction recently because of its applications in fields like visual entertainment, 3D reconstruction, autonomous driving, object detection [1], outdoor mapping, navigation and 3DTV [2], [3]. It is a research area that tries to imitate vision systems in humans by using two or several 2D views of the same scene to get three-dimensional depth details of the scene. It intends to find the corresponding relationship between matching pixels. A stereo matching algorithm uses stereo images that are rectified as an input [4], [5]. The horizontal displacement between the matching pixels is called disparity. With additional details, a disparity map could be transformed to a depth of scene. Disparity map accuracy is very crucial as small inaccuracies may affect the result. Obtaining an efficient and precise disparity map is a tedious task because of the existence of noise, occlusions, low textures, ill-posed regions, and the lighting conditions. Hence, it is significant to create a good disparity map.

Stereo matching techniques are classified as conventional algorithms and deep learning methods. Conventional algorithms are grouped into local and global algorithms. In local approaches, disparity is computed by comparing small areas [6] [7]. The disparity calculation relies on intensity in a defined support area. In real time the stereo images collected may be prone to noise, lighting distortions which reduces the efficiency of these algorithms. To overcome these drawbacks, a census transform in stereo matching is proposed in [8] which can decrease the effect of amplitude distortion. It aims at mapping the pixels to a binary string and then calculates the similarity between the pixels by means of Hamming distance. But, this method relies mainly on the central pixel, leading to false matching in a noisy environment. To reduce this shortcomings, a three-state census is proposed in [9] which is tolerant to any noise and enhances the robustness of stereo matching. An algorithm is implemented in [10] to perform census transform that reduces the noise interference and amplitude distortions in the images. A star-census transform (SCT) is introduced [11] that initiates the neighborhood pixel sampling in a symmetrical order that excludes the central pixel in the matching window. An improvised AD-Census stereo matching using gradient fusion (ADSG) is introduced in [12]. The absolute difference is used along with census transform for cost calculation, the result is then combined with gradient cost. These methods focus only on the information locally and hence have a low complexity and execute in shorter time. But the results generated by these local methods in the areas of occlusion, texture less and discontinuities is not satisfying.

The semi global algorithm was proposed in [13]. The accuracy and computational efficiency of semi global algorithms lies in between that of local and global algorithms. A global method considers disparity computation as a global energy minimization method for all disparity values. The energy function has two terms namely data term which penalizes pixels with inconsistent values and smoothness term with enforces smoothing constraint by considering the neighboring pixels. Some of the commonly used global algorithms are graph cuts algorithm [14] and belief propagation technique [15]. A disparity estimation based on tree structure named Pyramid-tree is introduced in [16]. It performs cross regional smoothing that can handle low texture regions. Global methods can generate a good quality disparity map, but they are also quite expensive and time-consuming.

Deep stereo methods are popular these days. Zbontar et al. [17] used a network to get patch-wise details to compute matching cost. The proposed network is trained to find the similarity that exists between a pair of images. It is then processed using classic post processing. The GC-Net [18] is a network with a high performance. It applied 3D convolution kernel to the correspondence space and proposed disparity refinement. This provided improvement over the previous approach. A pyramid stereo matching network is proposed in [19]. It improved the feature extraction by means of multi scale feature extraction network [20]. A network namely cascaded residual learning [21] was introduced which uses a DispNet. This is made up of two sub parts called DispFullNet and a DispResNet. The first network computes the raw disparity map. The second network tries to optimize the raw disparity map by computing the multiscale residual information. Williem et al. [22] introduced a method known as self-guided cost aggregation that uses a convolution network for local stereo matching. The network is made up of emotional weight network and descent filtering network. In LEA Stereo [23] a search is performed to streamline matching pipeline. Shivam Duggal et al. [24] developed a trainable network. Many recent papers introduced refinement components steps to improvise the disparity map quality. The MSMD-Net [25] introduced multi scale technique in which the stereo images are processed using multi resolution pyramid network. The RAFT- stereo [26] consists of a network for stereo estimation along with refinement stage. The deep learning-based methods can produce depth map from a given stereo image pairs, but these stereo methods still find it difficult to find correct correspondences in texture less and the occluded regions.

Though several techniques have been proposed to improvise the matching accuracy, the low accuracy in the occluded and texture less regions has not been handled very well. A depth estimation technique using improvised census transform and disparity refinement using deep learning to enhance the results in occluded and texture less regions is proposed . The weighted sum of the center pixel and its four neighbors is used to calculate the center pixel value in the improvised census transform in order to reduce noise in initial disparity map. The occluded and texture less regions of initial disparity map are refined using Generative adversarial network (GAN) deep learning framework. The extensive experiments performed on Middlebury datasets shows the efficacy of our method. Our method improvises the efficiency of disparity map by a considerable amount. The suggested method is explained in Section II. The outcomes of the suggested method are shown in Section III. In Section IV, the paper's conclusions are discussed.

II. Methodology

The proposed method applies improved census transform is applied for the stereo images and a matching cost is obtained using Hamming distance. Then, a cost aggregation is carried using semi global method to compute an initial disparity map. Finally, a disparity refinement network using GAN is proposed to increase the efficiency of disparity map from which depth is estimated. An overview of the whole methodology is Fig. 1.

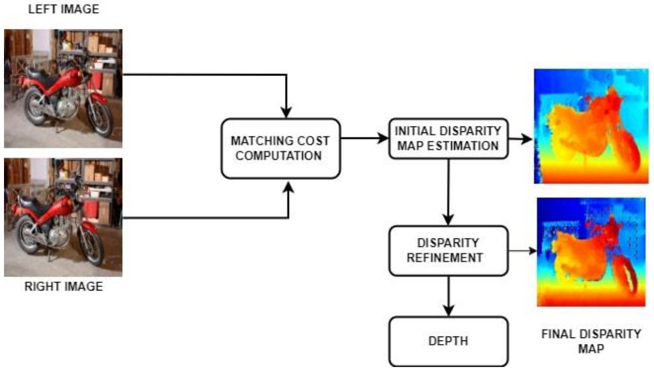

Fig. 1. Block diagram of the methodology.

A. Improves Census Transform

Methods for stereo matching based on intensity difference contain a lot of errors, especially for the outdoor images. To overcome these drawbacks Census transform (CT) method is used for computing matching cost. It is a local method that relies on relative ordering of pixels rather than intensity within a fixed window. Hence it can efficiently handle radiometric variations like lighting changes and illumination differences and discontinuities. The traditional CT is shown in Fig. 2. Census transform consider the center pixel value, compares with all the remaining pixels and assigns the 1 if the center pixel value is less than the compared pixel, otherwise 0 is assigned. It is then represented as a binary bit string.

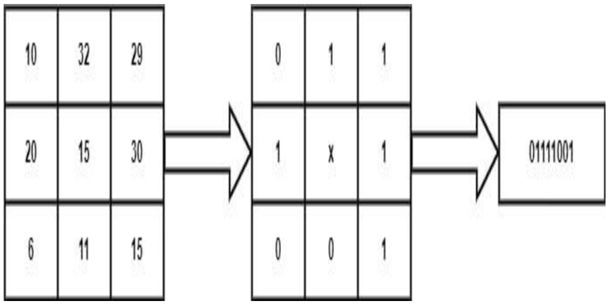

Fig. 2. Traditional census transform.

Census transform is represented by the following equation

$$T_{cen} = \otimes_{i\in w_p} \xi[I(p), I(m)] \quad (1)$$

$$\xi[I(p), I(m)] = \begin{cases} 1, & I(p) \leq I(m) \\ 0, & otherwise \end{cases} \quad (2)$$

Here, $\otimes$ is a bitwise operation, w_p represent window with p as centre pixel, m is any point in w_p and $I(p), I(m)$ are pixel values of points p and m respectively.

The traditional CT can reduce the impact of distortions in amplitude, but it depends heavily on the middle pixel. It is prone to noise, as it measures the relative difference of the neighboring pixels based on middle pixel. When the center pixel is affected by noise, encoding from the census transform might vary drastically which may lead to mismatched pixels. Due to noise if the center pixel value changes from 15 to 35, the traditional CT transformation for a 3 X 3 patch of image is depicted in Fig. 3. Since the traditional CT depends on the pixel at the center, the noisy center pixel value 35 is considered to compare with the remaining pixels in the image patch. The census code obtained is 00000000. Here there is a difference in 5 bits as compared to the initial code 01111001.

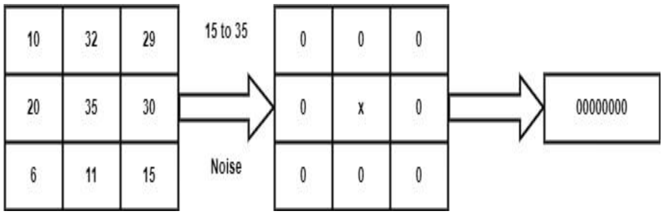

Fig. 3. Traditional census transform in noisy condition.

Aiming to overcome this drawback, an improvised census method is proposed in this paper where the weighted summation of pixel at the center and the four neighboring pixels is used to update the center pixel.

Let $I(m,n)$ be the center pixel. The weight distribution of pixel of the center and the four neighboring pixels are:

$$\mathrm{wt(m,n)} = 0.4 \qquad (3)$$

$$\mathrm{wt(m+1,\ n)} = 0.15 \qquad (4)$$

$$\mathrm{wt(x-1,\ y)} = 0.15 \qquad (5)$$

$$\mathrm{wt(m,n+1)} = 0.15 \qquad (6)$$

$$\mathrm{wt(m,\ n-1)} = 0.15 \qquad (7)$$

The weights are assigned in such a manner that the weight of each pixel lies in between 0 and 1 and the total weighted sum of the pixel at the center and four neighboring pixel is 1.

The weighted sum of pixel centered at (x, y) is computed using the following equation.

$$I_{wt}(m,n) = I(m,n)wt(m,n) + I(m+1,n)wt(m+1,n) + I(m-1,n)wt(m-1,n) + I(m,n+1)wt(m,n+1) + I(m,n-1)wt(m,n-1) \qquad (8)$$

The following equation is used to update the value of the center pixel.

$$I_{mid}(m,n) = \begin{cases} \{\ I(m,n), & |I_{wt}(m,n) - I(m,n) \le\ T| \\ I_{wt}(m,n), & |I_{wt}(m,n) - I(m,n) >\ T| \end{cases} \qquad (9)$$

If the variation between the weighted sum of center pixel and the original center pixel is more than the threshold $T = 6$, then the pixel value in the center is updated by the weighted sum otherwise the original is used.

The improvised technique for census transform proposed in the paper is depicted in Fig. 4.

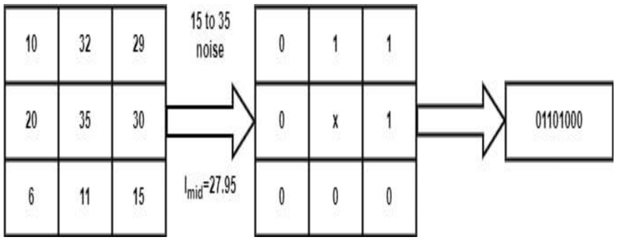

Fig. 4. Census transform of the proposed method in noisy condition.

Due to noise if center pixel value change from 15 to 35, the improvised method proposed in the paper updates the value of center pixel to 27.95 using the weighted sum as shown in Fig. 4 (i.e 35*0.4+11*0.15+32*0.15+30* 0.15+20*0.15). This value is compared with the remaining pixels in the patch to get the census code 01101000. The pixels differ only by 2 bit as compared to original code 01111001. This demonstrates how the approach is noise-resistant and improves matching performance.

B. Matching Cost Computation

To ascertain whether the values between two pixels indicate the matching point of a scene, a matching computation of cost is carried out. After the census transform the correspondence of pixels can be determined using Hamming distance [8]. The Hamming distance between matching points is found to estimate the correspondence between matching points. Let $S_{cenL}(p)$ be a binary bit array of pixel p in the left stereo image and $S_{cenR}\ (q)$ be the binary bit array of pixel q in right stereo image for disparity d . The following calculation uses the Hamming distance to compute the matching census cost between p and q.

$$C(p,d) = Hamming[S_{cenL}(p), S_{cenR}\ (q)] \qquad (10)$$

The cost computation for the center pixel of $3\ X\ 3$ image patch is depicted in Fig. 5.

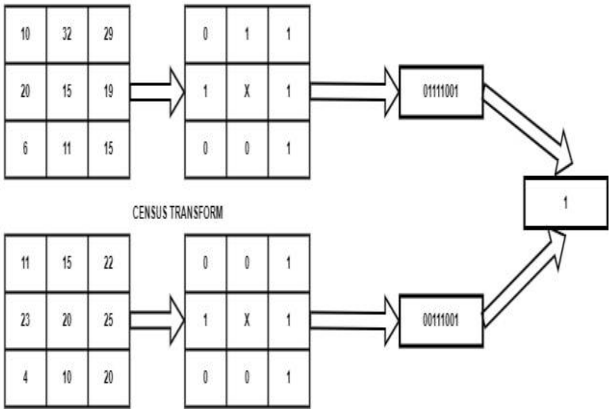

Fig. 5. Matching cost computation.

C. Initial Disparity Estimation

Due to noise, the pixel wise cost may produce ambiguous results. Hence additional constraint is included to get a smooth disparity by penalizing the changes in the neighboring pixels [27]. The smoothness constraint and pixel wise cost is represented by the energy function $E(D)$.

$$E\ (D) = \sum_P C(p, D_P\ ) + \sum_{x\in N_p} P_1\ T\big[\big|D_p\ -\ D_x\big| = 1\big] + \sum_{x\in N_p} P_2\, T\big[\big|D_p\ -\ D_x\big| > 1\big] \qquad (11)$$

$C(p, D_P\ )$ is the cost summation of all the pixel for disparity D. P_1 is the penalty applied to pixels x in N_P with low disparity difference. P_2 is the penalty for pixels in N_p with high disparity difference.

The stereo matching problem aims to minimize the energy function$E(D)$. Finding the minimum energy function $E(D)$ is computationally expensive. The energy function is approximated by aggregating the matching cost from all directions $'r'$. The total number of directions r is 8. The cost at direction r is represented by,

$$S_r\ (p,d) = C(p,d) + min\left[S_r(p-r,d), S_r\ (p-r,d-1) + P_{1,}\ S_r(p-r,d+1), S_r(p-r,i) + \ P_2\ \right] \quad (12)$$

$C(p,d)$ is cost for pixel p and $S_r(p-r,d)$ is the pixel cost at direction r with disparity d, $S_r\ (p-r,d-1)$ is cost at direction $'r'$ and disparity $'d-1'$. $S_r(p-r,d+1)$ denotes cost for disparity $'d+1'$ and direction $'r'$. S_r is minimum pixel cost at direction r.

The following equation is then used to determine the initial disparity.

$$D(p) = argmin \sum S_r\ (p,d) \quad (13)$$

D. Disparity Refinement

The initial disparity calculated may contain wrongly matched disparities at the object boundaries, occluded areas and the texture less regions. Finding the correct disparities in these areas is a difficult task. Hence, an appropriate disparity refinement method is needed. A disparity refinement is performed based on deep learning-based technique using the generative adversarial network (GAN) to handle texture less, occluded, and discontinuous areas. The method proposed uses GAN network introduced by Good fellow [28] for disparity refinement. GAN includes two networks called generator and a discriminative network that are implemented based on neural networks. The generator takes initial disparity map as its input and focuses on generating a refined disparity map. The discriminator is fed with ground truth disparity along with disparity map produced by the generator. The discriminator aims to differentiate the ground truth disparity and generated refined disparity map. The feedback from the discriminator is given to the generator to fine tune the generated image. This procedure is repeated until the resulting disparity resembles the ground truth disparity. The disparity refinement network proposed in the paper is depicted in Fig. 6.

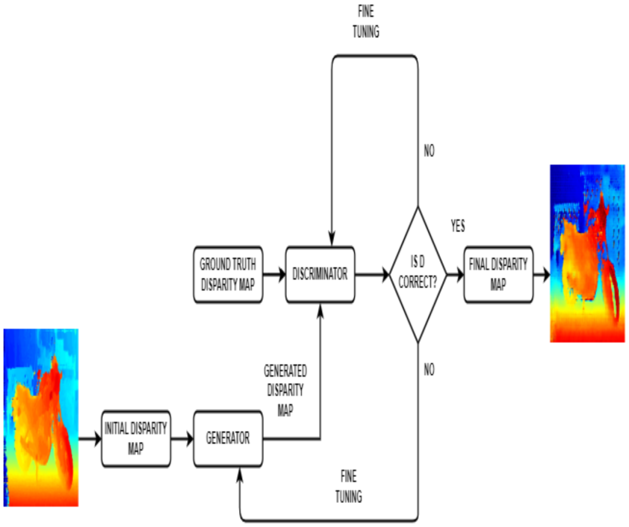

Fig. 6. Architecture of disparity refinement network.

A Pix2Pix GAN [29] is used to refine the disparity map. Pix2Pix GAN is an adversarial network. Pix2Pix GAN is known for the capacity of producing high quality images. The initial disparity is given as input to the generator. The generator generates the disparity map which is then fed to the discriminator. The various generator networks available are UNET 128, ResNet 6 and Resnet 9. The proposed disparity refinement network uses UNET128 as it can learn with few training images. The architecture of UNET 128 generator used in the proposed approach is depicted in Fig. 7.

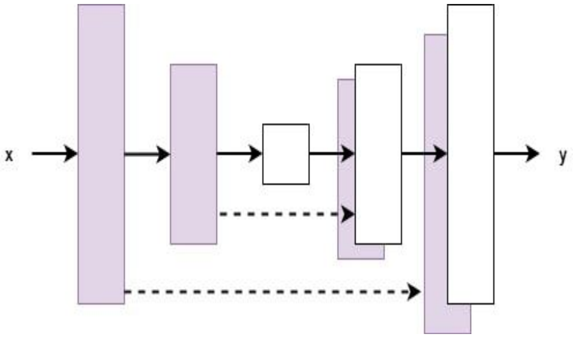

Fig. 7. Architecture of generator.

It uses a network consisting of several convolutional layer, batch normalization, dropout, and activation layers. It is trained using adversarial loss and then revised by means of L_1 loss. This loss drives the generator to generate image close to ground truth disparity. The generator is then updated using a sum of L_1 loss and a loss called as adversarial loss. A comparative study of UNet 128 with other networks such as ResNet 6 and ResNet 9 is given in Table I. ResNet 6 and Resnet 9 are the deep residual networks which include 6 residual blocks and nine residual blocks respectively. The information details are passed via a shortcut connection. Convolutional, batch-normalization, and corrected Liner Unit (ReLU) layers make up a traditional residual block. For evaluation, measurements like squared relative difference (SRD) and absolute relative distance (ARD) are used. Lower values of the above metrics indicate better performance. The efficiency of the generator architecture is shown in Table I. Results obtained for UNet 128 is better than the ResNet 6 and ResNet 9.

$$ARD = \frac{1}{N}\sum \frac{d_t\,(p,q) - d_g(p,q)}{d_t(p,q)} \quad (14)$$

$$SRD = \frac{1}{N}\sum \frac{|d_t(p,q) - d_g(p,q)|^2}{d_t(p,q)} \quad (15)$$

Here d_t is generated disparity map, d_g is ground truth disparity map. N is the total pixels.

TABLE I. COMPARATIVE STUDY OF GENERATOR ARCHITECTURE

	ResNet 6	ResNet 9	UNet 128	
ARD	0.058	0.051	0.037	Lower is better
SRD	0.409	0.413	0.403	

The discriminator is based on Patch GAN model. This PatchGAN model provides extremely high frequency information. The GAN's primary objective is described as,

$$L_{GAN}(G,D) = E_{p,q}\left[logD(p,q)\right] + E_{p,r}\left[log\left(1 - D\left(p, G(p,r)\right)\right)\right] \quad (16)$$

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

The generator G attempts to decrease the objective as response to the discriminator D which attempts to increase it. The result is as follows:

$$G^{*} = argmin_{g} max_{d} L_{GAN}(G,D) \quad (17)$$

The aim of G is to decrease the objective and the generator updates itself using Loss L_1. It is computed as,

$$Loss_{L_1}(G) = E_{p,q,r}\left[\|\left(q - G(p,r)\right)\|_1\right] \quad (18)$$

The objective is updated as

$$G^{*} = argmin_{g} max_{d} L_{GAN}(G,D) + \lambda Loss_{L_1}(G) \quad (19)$$

The performance of training model is depicted in Fig. 8. Here the training loss decreases gradually as the number of epochs increases. Lower the loss, the more effective the model is.

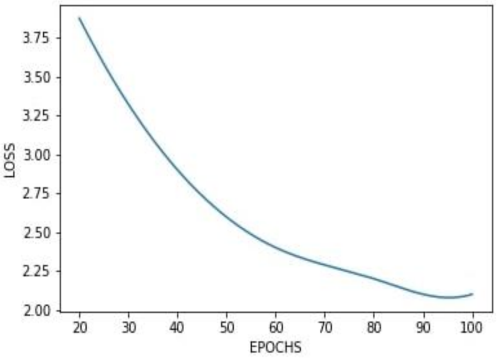

Fig. 8. Training loss versus epochs.

E. Depth

Once the disparity map is generated for the stereo images, the depth Z is estimated using the following formula:

$$Z = \frac{f\ X\ B}{D} \quad (20)$$

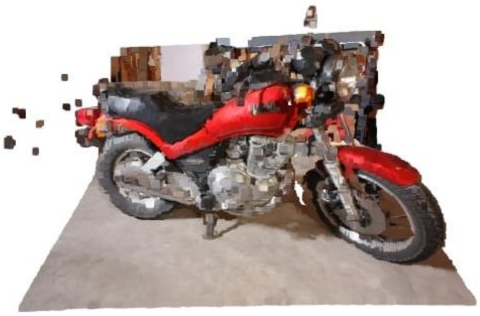

Fig. 9. Point cloud generated for motorcycle image.

Here f is focal length, B is stereo camera baseline. These values are obtained from the stereo calibration. Once the depth is estimated the exact coordinates of each pixel in the scene can be computed. These coordinates are made used to construct point clouds. The point cloud generated for Motorcycle image is shown in Fig. 9. The coordinates are stored in polygon format file. The output ply file is plotted using ply file plotters. The point cloud shown above was constructed using open3d.

III. Results and Discussion

The experiments were performed using Middlebury dataset [30], [31] images to analyze the performance. The refinement network is trained using the Pytorch framework on a personal machine. The computer hardware environment used is a Dual Intel-Xeon E5-2609V4 8C having 1.7 GHz 20M 6.4 GT/s and 128GB Memory. A Dual NVDIA Tesla server P100 GPU having 3584Cores and maximum of 18.7 TeraFLOPS is used. The datasets are downscaled to 256 pixels width and 256 pixels height for computational purposes. The Adam optimizer is used to optimize the discriminator. The learning rate is 0.0002. The GAN models does not converge, hence a balance has to be established between the generator and discriminator. The number of epochs is 100.

A. Middlebury Dataset

The Middlebury dataset includes rectified stereo images from indoor and outdoor surroundings utilizing a stereo vision concept. These images are complex, and it has images of different characteristics such as, different resolutions and low texture areas. Hence, the dataset consists of complex images for framework evaluation. Our stereo matching technique is robust to occluded and non-textured regions. The details of testing images like Cones, Teddy and Venus from Middlebury 2001 and 2003 are given in Table II. The details of higher resolution images from Middlebury 2014 are given in Table III.

TABLE II. IMAGE OF MIDDLEBURY 2001 AND 2003

Images	Disparity level	Image resolutions
Cones	60	450 × 375
Teddy	60	450 × 375
Venus	20	434 × 383

TABLE III. IMAGE OF MIDDLEBURY 2014

Images	Disparity level	Image resolutions
Adirondak	73	718 × 496
ArtL	64	347 × 277
Jadeplant	160	659 × 497
Motorcycle	70	741 × 497
Pipes	75	735 × 485
Playroom	83	699 × 476
Playtable	73	680 × 463
PlaytableP	73	681 × 462
Recycle	65	720 × 486
Shelves	60	738×497

B. Noise Resistance Test

The traditional CT heavily depends on the center pixel. If this pixel is prone to noisy condition, then the census encoding may differ which leads to mismatches. To analyze the efficiency of the proposed improved census transform in a noisy condition, salt and pepper noise of 2% and 5% noise is applied to Cones, Teddy and Venus images. The qualitative results for Teddy image when 2% salt and pepper noise is applied is represented in Fig. 10 and Table IV shows the percentage of bad matching pixels (PBMP) of the initial disparity map for the proposed improved census transform and traditional CT. The outcome conclude that results of the method proposed in the noisy condition is remarkably good than the traditional CT.

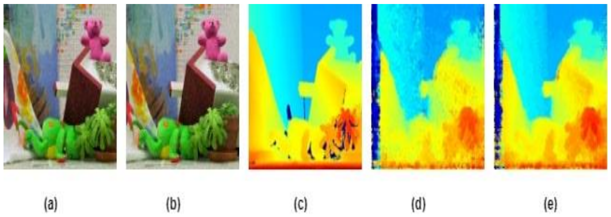

Fig. 10. Visual results for initial disparity map on noisy Teddy image (a) Left reference image (b) Right image (c) Ground Truth Disparity (d) Initial Disparity Map using traditional CT (e) Initial Disparity Map using proposed method.

TABLE IV. PBMP OF INITIAL DISPARITY MAP

	Salt & Pepper Noise (2%)		**Salt & Pepper Noise (5%)**	
	Traditional CT	**Proposed**	**Traditional CT**	**Proposed**
Cones	8.984	7.556	12.990	8.980
Teddy	13.492	9.715	19.492	12.103
Venus	3.229	1.370	6.062	2.740

C. Qualitative Results

The initial and improved disparity maps estimated by the suggested method are shown in Fig. 11. In the Fig. 11, the Fig. 11(a) is the left reference image. Fig. 11(b) is the right image. The ground truth disparity is given Fig. 11(c). The fourth column Fig. 11(d) represent initial disparity map. The refined disparity map is represented in Fig. 11(e). The red rectangular regions marked in Fig. 11(d) represent the occluded areas which are filled in the refined disparity map obtained by the suggested approach. The yellow circular region marked in the initial disparity of the Venus image shows the texture less region which is filled in the refined disparity map. It is discovered that the suggested method effectively creates high-quality disparity maps in noisy, textureless, and occluded regions.

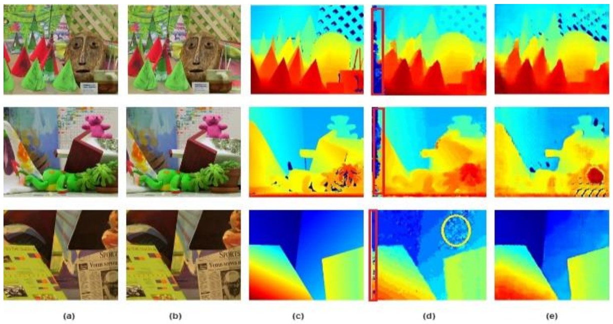

Fig. 11. Visual results on Cone, Teddy, and Venus images (a) Left image (b) Right image (c) Ground Truth Disparity (d) Initial Disparity Map (e) Refined disparity map.

The disparity maps generated for images such as Jade Plant, Adirondack, Motorcycle and Recycle are presented in first, second, third and fourth rows respectively in Fig. 12. The first row of the Fig. 12 shows a Jade Plant image from Middlebury dataset. This image is very challenging to match due to brightness difference. But, the proposed method has correctly discovered the disparities. The second and third rows of Fig. 12 shows Adirondack and Motorcycle images. The texture less surfaces in the initial disparity map of Adirondack image is highlighted by the yellow circular region. These regions are well recreated by the proposed approach. The fourth row of Fig. 12 shows Recycle image. The occluded areas in the initial disparity map is highlighted by the red rectangular region. The possibility of getting wrong matches in these regions are very high. These occluded areas are filled accurately in the estimated disparity map. We find that the proposed method produces efficient results in occluded and texture-less regions.

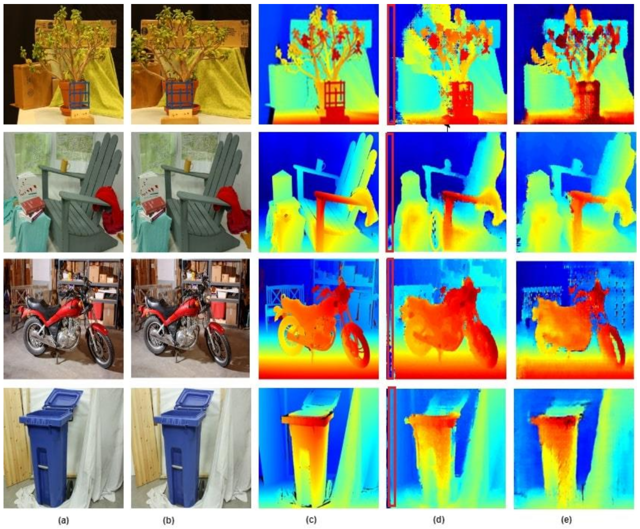

Fig. 12. Visual results on Jade Plant ,Adirondack, , Motorcycle and Recycle images (a) Left reference image (b) Right image (c) Ground Truth image (d) Initial Disparity Map (e) Refined disparity map.

D. Evaluation Metrics

The quantitative analysis is performed using the evaluation metrics namely root mean square error (RMSE) and PBMP. The efficiency increases when PBMP and RMSE values decrease. N be the number of pixels. d_t and d_g be the disparity map estimated and ground truth disparity maps respectively.

RMSE is calculated as:

$$RMSE = \left[\frac{1}{N}\sum\left|d_t(x,y) - d_g(x,y)\right|^2\right]^{\frac{1}{2}} \quad (21)$$

PBMP is calculated as follows:

$$PBMP = \left[\frac{1}{N}\sum\left|d_t(x,y) - d_g(x,y)\right| > T\right] * 100 \quad (22)$$

E. Comparison with Existing Methods

The proposed method is compared with ADSG [12] and Deep Pruner [24]. The results for the methods compared are obtained from Middlebury evaluation leader board. An improved AD-Census method using gradient fusion is used in

ADSG. The absolute difference and census transform is used for cost calculation, which is then combined with gradient cost. This method focus only on local information and hence do not give satisfying results in the areas of occlusion, texture-less and discontinuities. Deep Pruner uses a trainable network. It do not produce satisfactory results in the occluded regions. Tables V and VI demonstrate that the comparison of RMSE and PBMP results.

TABLE V. COMPARISON OF RMSE RESULTS

Images	ADSG	Deep Pruner	Proposed
Adirondack	19.5	6.18	5.14
ArtL	24.6	9.50	6.22
Jade plant	25.8	28.2	5.44
Motorcycle	79.6	10.3	6.52
Pipes	32.1	13.9	6.74
Playroom	35.2	8.91	7.85
Playtable	50.0	4.89	3.52
PlaytableP	19.9	4.74	3.54
Recycle	17.6	3.81	5.55
Shelves	21.9	4.28	5.66
Avg	32.62	9.471	5.618

TABLE VI. COMPARISON OF PBMP FOR THRESHOLD=1

Images	ADSG	Deep Pruner	Proposed
Adirondack	38.9	39.7	23.97
ArtL	35.5	41.8	42.36
Jade plant	49.8	62.8	44.50
Motorcycle	43.2	45.3	45.92
Pipes	41.5	53.8	34.78
Playroom	57.8	57.7	26.08
Playtable	64.4	48.2	46.28
PlaytableP	42.2	41.7	47.15
Recycle	37.5	36.8	27.36
Shelves	65.0	54.2	45.20
Avg	47.58	48.2	38.36

Our technique yields the lowest average RMSE and PBMP, as shown in Table V and Table VI. This signifies the accuracy and competitiveness of our method as compared to ADSG [12] and Deep Pruner [24]. The proposed method produces average RMSE 5.68 and PBMP 38.36%. The improved census transform in matching cost is robust to noise. Additionally the disparity refinement based on deep learning-based technique using the generative adversarial network (GAN) handle texture less, occluded, and discontinuous areas and produce a good quality disparity map.

IV. CONCLUSION

A stereo matching method that is based on improvised census transform along with an optimization framework is proposed to determine the initial disparity map. Further disparity refinement is carried out using GAN to obtain the depth of a scene. The traditional census transform heavily depends on center pixel. If this pixel is prone to noise, then census encoding generated will differ which may lead to false matching. To handle this issue an improved census cost that relies on the weighted sum values is proposed. In the disparity refinement stage a deep learning based network using GAN is proposed which can handle outliers and enhance the correctness of matching. The efficiency of the suggested strategy is assessed using images from Middlebury benchmark. The comparison with the current system showed that the proposed method works better than other methods.

REFERENCES

[1] H. M. Wang, H. Y. Lin, and C. C. Chang, “Object detection and depth estimation approach based on deep convolutional neural networks,” *Sensors*, vol. 21, no. 14, 2021, doi: 10.3390/s21144755.

[2] M. Menze, C. Heipke, and A. Geiger, “Object Scene Flow,” *ISPRS J. Photogramm. Remote Sens.*, vol. 140, pp. 60–76, 2018, doi: 10.1016/j.isprsjprs.2017.09.013.

[3] S. Hong, M. Li, M. Liao, and P. Van Beek, “Real-time mobile robot navigation based on stereo vision and low-cost GPS,” *IS T Int. Symp. Electron. Imaging Sci. Technol.*, pp. 10–15, 2017, doi: 10.2352/ISSN.2470-1173.2017.9.IRIACV-259.

[4] R. A. Hamzah and H. Ibrahim, “Literature survey on stereo vision disparity map algorithms,” *J. Sensors*, vol. 2016, 2016, doi: 10.1155/2016/8742920.

[5] K. Y. Kok and P. Rajendran, “A review on stereo vision algorithms: Challenges and solutions,” *ECTI Trans. Comput. Inf. Technol.*, vol. 13, no. 2, pp. 134–151, 2019, doi: 10.37936/ecti-cit.2019132.194324.

[6] 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 J. Image Video Process.*, vol. 2020, no. 1, pp. 1–9, 2020, doi: 10.1186/s13640-020-00525-3.

[7] Deepa and K. Jyothi, “A Robust Disparity Map Estimation for Handling Outliers in Stereo Images,” *2021 5th Int. Conf. Electr. Electron. Commun. Comput. Technol. Optim. Tech. ICEECCOT 2021 - Proc.*, no. December, pp. 38–43, 2021, doi: 10.1109/ICEECCOT52851.2021.9708034.

[8] R. Zabih and J. Woodfill, “Non parametric local transforms.pdf,” 1994.

[9] Y. Men, N. Ma, G. Zhang, X. Li, C. Men, and P. Sun, “A stereo matching algorithm based on Census transform and improved dynamic programming,” *Harbin Gongye Daxue Xuebao/Journal Harbin Inst. Technol.*, vol. 47, no. 3, pp. 60–65, 2015, doi: 10.11918/j.issn.0367-6234.2015.03.010.

[10] N. Y. C. Chang, T. H. Tsai, B. H. Hsu, Y. C. Chen, and T. S. Chang, “Algorithm and architecture of disparity estimation with mini-census adaptive support weight,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 20, no. 6, pp. 792–805, 2010, doi: 10.1109/TCSVT.2010.2045814.

[11] J. Lee, D. Jun, C. Eem, and H. Hong, “Improved census transform for noise robust stereo matching,” *Opt. Eng.*, vol. 55, no. 6, p. 063107, 2016, doi: 10.1117/1.oe.55.6.063107.

[12] H. Liu *et al.*, “Stereo matching algorithm based on two-phase adaptive optimization of AD-census and gradient fusion,” *2021 IEEE Int. Conf. Real-Time Comput. Robot. RCAR 2021*, pp. 726–731, 2021, doi: 10.1109/RCAR52367.2021.9517511.

[13] H. Hirschmüller, “Accurate and efficient stereo processing by semi-global matching and mutua information,” *Proc. - 2005 IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognition, CVPR 2005*, vol. II, no. 2, pp. 807–814, 2005, doi: 10.1109/CVPR.2005.56.

[14] V. Kolmogorov and R. Zabih, “Computing visual correspondence with occlusions using graph cuts,” *Proc. IEEE Int. Conf. Comput. Vis.*, vol. 2, pp. 508–515, 2001, doi: 10.1109/iccv.2001.937668.

[15] J. Sun, H. Y. Shum, and N. N. Zheng, “Stereo matching using belief propagation,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 2351, no. 7, pp. 510–524, 2002, doi: 10.1007/3-540-47967-8_34.

[16] 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 Trans. Circuits Syst. Video Technol.*, vol. 31, no. 10, pp. 4007–4019, 2021, doi: 10.1109/TCSVT.2020.3044891.

[17] J. Žbontar and Y. Lecun, “Stereo matching by training a convolutional neural network to compare image patches,” *J. Mach. Learn. Res.*, vol.

17, pp. 1–32, 2016.

[18] A. Kendall *et al.*, “End-to-End Learning of Geometry and Context for Deep Stereo Regression,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 66–75. doi: 10.1109/ICCV.2017.17.

[19] R. Qin, X. Huang, W. Liu, and C. Xiao, “Pairwise stereo image disparity and semantics estimation with the combination of U-net and pyramid stereo matching network,” *Int. Geosci. Remote Sens. Symp.*, vol. 2019-July, pp. 4971–4974, 2019, doi: 10.1109/IGARSS.2019.8900262.

[20] X. Li, T. Lai, S. Wang, Q. Chen, C. Yang, and R. Chen, “Weighted feature pyramid networks for object detection,” *Proc. - 2019 IEEE Intl Conf Parallel Distrib. Process. with Appl. Big Data Cloud Comput. Sustain. Comput. Commun. Soc. Comput. Networking, ISPA/BDCloud/SustainCom/SocialCom 2019*, pp. 1500–1504, 2019, doi: 10.1109/ISPA-BDCloud-SustainCom-SocialCom48970.2019.00217.

[21] 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,” *Proc. - 2017 IEEE Int. Conf. Comput. Vis. Work. ICCVW 2017*, vol. 2018-Janua, pp. 878–886, 2017, doi: 10.1109/ICCVW.2017.108.

[22] Williem and I. K. Park, “Deep self-guided cost aggregation for stereo matching,” *Pattern Recognit. Lett.*, vol. 112, pp. 168–175, 2018, doi: 10.1016/j.patrec.2018.07.010.

[23] X. Cheng *et al.*, “Hierarchical neural architecture search for deep stereo matching,” *Adv. Neural Inf. Process. Syst.*, vol. 2020-Decem, no. NeurIPS, pp. 1–12, 2020.

[24] S. Duggal, S. Wang, W. C. Ma, R. Hu, and R. Urtasun, “Deeppruner: Learning efficient stereo matching via differentiable patchmatch,” *Proc. IEEE Int. Conf. Comput. Vis.*, vol. 2019-Octob, pp. 4383–4392, 2019, doi: 10.1109/ICCV.2019.00448.

[25] Z. Shen, Y. Dai, and Z. Rao, “MSMD-Net: Deep Stereo Matching with Multi-scale and Multi-dimension Cost Volume,” 2020, [Online]. Available: http://arxiv.org/abs/2006.12797

[26] L. Lipson, Z. Teed, and J. Deng, “RAFT-Stereo: Multilevel Recurrent Field Transforms for Stereo Matching,” *Proc. - 2021 Int. Conf. 3D Vision, 3DV 2021*, pp. 218–227, 2021, doi: 10.1109/3DV53792.2021.00032.

[27] H. Hirschmüller, “Stereo processing by semiglobal matching and mutual information,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 2, pp. 328–341, 2008, doi: 10.1109/TPAMI.2007.1166.

[28] I. Goodfellow *et al.*, “Generative adversarial networks,” *Commun. ACM*, vol. 63, no. 11, pp. 139–144, 2020, doi: 10.1145/3422622.

[29] P. Isola, J. Y. Zhu, T. Zhou, and A. A. Efros, “Image-to-image translation with conditional adversarial networks,” in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017, pp. 5967–5976. doi: 10.1109/CVPR.2017.632.

[30] D. Scharstein and R. Szeliski, “High-accuracy stereo depth maps using structured light,” *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 1, no. June, pp. 195–202, 2003, doi: 10.1109/cvpr.2003.1211354.

[31] X. Jiang, J. Hornegger, and R. Koch, “Pattern recognition: 36th German Conference, GCPR 2014 Münster, Germany, September 2–5, 2014 proceedings,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 8753, no. 2, pp. 31–42, 2014, doi: 10.1007/978-3-319-11752-2.