ENHANCED DEPTH MAP IN LOW LIGHT CONDITIONS FOR STEREO RGB CAMERAS

Joseph Chang, Truong Nguyen

Department of Electrical and Computer Engineering, University of California, San Diego

ABSTRACT

Most existing depth estimation methods deal with daytime scenarios. However, models trained for daytime scene do not perform well in low-light situations due to the lack of clear environmental features, glare, overexposure from lights, and noisy images. This is problematic for safe autonomous driving as pedestrian and guardrail detection at night is very challenging and poses potential life-threatening situations. This paper addresses this issue by improving the quality of disparity maps obtained in low-light based on previous work. We introduce an algorithm based on defogging methods which first preprocesses night images to improve their luminance. Then, we train a generative adversarial network which iteratively learns the correct disparity prediction. The experiments verify the improved performance of the proposed method.

***Index Terms*—** computer vision, depth estimation, low light, generative adversarial networks

1. INTRODUCTION

Depth Estimation based on stereo vision is an important topic in computer vision. It is used to gain a 3D understanding of the world from 2D images [1][2]. Although monocular depth estimation has lower performance compared to stereo-based depth estimation, it may enable many 3D datasets in the future since the depth of any image can be computed [3][4]. Another popular method is to use Lidar to form a depth map by measuring a laser's return time from its target. Although Lidar-based methods are highly accurate, they require expensive equipment for longer-range applications such as driving scenarios [5]. Hence, stereo RGB camera are currently the best method for depth estimation as it produces the most accurate depth map and does not require additional sensors. Despite the importance of depth estimation to existing applications such as virtual reality, autonomous cars, and robotic surgery, the low-light and nighttime scenarios have hardly been explored which impedes autonomous driving capability.

In this paper, our goal is to create an algorithm to improve the quality of depth maps in low-light conditions for autonomous driving using unsupervised learning. There are few nighttime 3D datasets for cars with complete ground truth available which makes supervised learning unsuitable. However, Oxford RobotCar provides nighttime car datasets with sparse depth from Lidar. Hence, our method utilizes this dataset to train a model that accounts for nighttime features and will accurately generate a disparity map given any nighttime RGB image. We create a collection of data that mimics various real-world environments by including glare, overexposed light, and poorly lit pedestrians. The model is tested with various nighttime images using meaningful evaluation metrics and shown to outperform current state-of-the-art methods in stereo nighttime vision. In the future, the improved depth map from the proposed algorithm will be used for object detection and classification in nighttime autonomous driving scenario.

2. METHOD

Given a RGB stereo image pair of a nighttime scene, our goal is to calculate the disparity map where each pixel represents the distance from the stereo camera to a point in space. The approach has two parts. First, it uses a defogging algorithm to improve the image quality, then it uses a neural based GAN approach to find the disparity.

Dataset: We use the Oxford RobotCar Dataset [6][7] as it contains a mixture of daytime and nighttime streetview data in various weather conditions. The images are stored in a raw format which requires conversion to RGB using a lookup table (LUT). The data is then ready for preprocessing and training. However, we need sparse ground truth disparity to calculate error. This requires the camera model of the Bumblebee XB3, 2D Lidar data from the front bumper, INS data, and timestamps from the Lidar and stereo camera for each dataset. The Lidar is projected into the camera, resulting in a depth map which is converted to a disparity map. The complete process is shown in Fig. 1.

Preprocessing: Given a nighttime image I, we flip its luma channel using 255-I and realize it looks like a foggy daytime image as shown in Fig. 2. We utilize this idea to construct a preprocessing procedure that takes nighttime images and enhances them through the defogging process. The improved contrast makes the images appear as if they were taken during twilight where objects have clearer edges and texture. This will give the network more features to extract while training and lead to better disparity prediction on nighttime images. In our case, the training and testing datasets are defogged us-

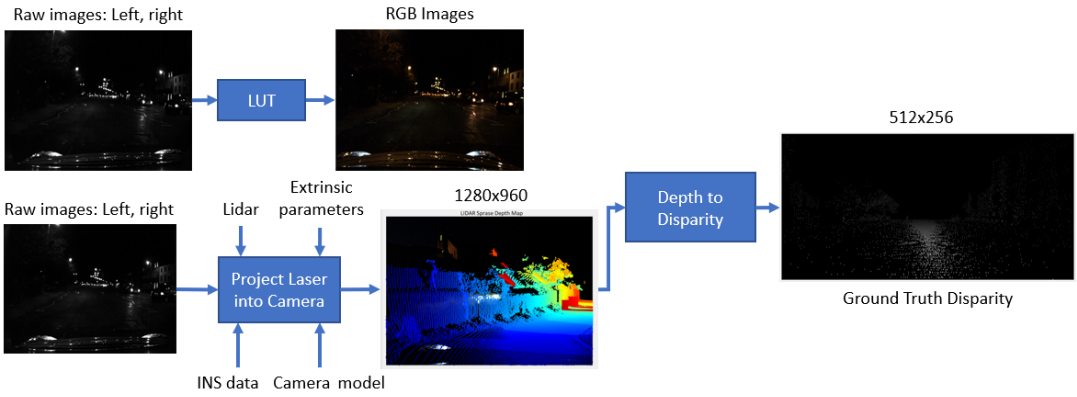


Fig. 1: Dataset Processing for Color Conversion and Ground Truth

ing a joint contrast enhancement and turbulence mitigation method (CETM) [8] designed to mitigate turbulence and remove fog in images while being fast enough for real-time applications. First, we convert the nighttime images to the YUV domain and then invert the luma channel which contains the brightness or luminance. CETM is then applied to denoise and defog the images, with atmospheric blur reduction. The result is smoothed by feature tracking using optical flow. To restore the image, we take 255-I on the luma channel and then convert the image back to RGB.

Fig. 2: Inverted Night Image

Fig. 3: Preprocessing Defog Step

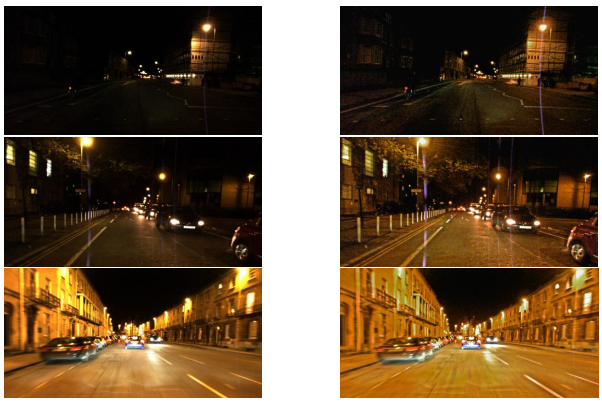


Fig. 4: Oxford Data with Defog Enhancement

Training: To train the depth estimation model, we utilize a GAN [9] introduced by Sharma et al. [10]. We begin by pretraining weights for daytime depth estimation using stereo daytime images X_{LD},X_{RD} and their corresponding ground truth Y_{LD}, as shown in Fig. 5. The images are passed through a stacked hourglass CNN model for training which gives the daytime disparity estimate weights f_D.

Next, we run the images through two separate training cycles. Each training cycle consists of a translation network [11] and a stereo network modeled after PSMNet [12]. In the first cycle, stereo daytime images X_{LD},X_{RD} shown in Fig. 6 are passed through a daytime generator g_D that renders an estimate of corresponding stereo nighttime images $\hat{X}_{LN}$,$\hat{X}_{RN}$. X_{LD},X_{RD} and $\hat{X}_{LN}$,$\hat{X}_{RN}$ are passed through the pretrained daytime weights f_D and nighttime weights f_N initialized as f_D, respectively. Stereo-consistency then computes the similarity of the two disparity estimates and updates f_N and g_D with its discriminator d_D. In addition, $\hat{X}_{LN}$,$\hat{X}_{RN}$ are passed through a nighttime generator g_N which renders stereo daytime images X_{LD},X_{RD} for the second training cycle.

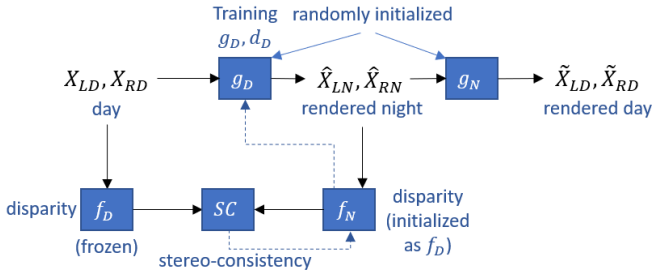
In the second training cycle, stereo nighttime images X_{LN},X_{RN} shown in Fig. 6 are passed through a nighttime generator g_N which renders corresponding daytime images $\hat{X}_{LD}$,$\hat{X}_{RD}$. $\hat{X}_{LD}$,$\hat{X}_{RD}$ is passed through a daytime generator to produce rendered nighttime images $\tilde{X}_{LN}$,$\tilde{X}_{RN}$. $\hat{X}_{LD}$,$\hat{X}_{RD}$ and $\tilde{X}_{LN}$,$\tilde{X}_{RN}$ are passed through the f_D from pretraining and f_N from the first cycle, respectively. Then, stereo-consistency compares the similarity of the disparity estimates and updates the weights of g_N and its discriminator d_N. This process iterates for as many epochs as specified until weights g_N^*, g_D^*, and f_N^* are optimized. To test nighttime images X_{LN},X_{RN}, we pass them through g_N^*, g_D^*, and f_N^* resulting in estimated disparity y_{LN}^*.

3. IMPLEMENTATION

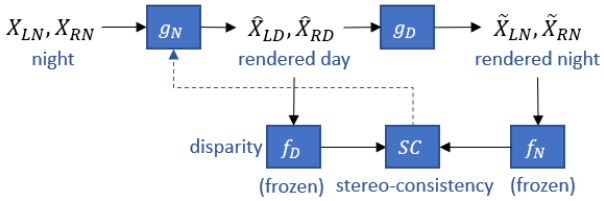
Preprocessing: Our algorithm requires initial preprocessing of nighttime data. We prepare our dataset with images from

the Oxford RobotCar Dataset [6][7]. For the training dataset, we randomly select 9000 stereo daytime images from the 2014-07-14 dataset. Additionally, we randomly select 5925 nighttime images from 2014-11-14 and 2975 from 2014-12-10. For our validation dataset, we randomly select 1000 nighttime images from 2014-11-14. All nighttime data is enhanced using the CETM algorithm described previously fto remove fog, increase contrast, and brighten the images. For validation nighttime data specifically, we obtain ground truth disparity by projecting Lidar scans into the camera [13].

Fig. 5: Pretraining Daytime Image Model f_D



(a) First Training Cycle



(b) Second Training Cycle

Fig. 6: GAN Training Cycle for Daytime and Nighttime

Network Architecture: Each translation network is a GAN containing a daytime and nighttime generator g_D and g_N to predict disparity. Each generator contains an encoder with three convolution layers and four residual blocks [14] followed by a decoder with four residual blocks, two deconvolution layers, and one convolution layer. Each translation network also has daytime and nighttime discriminators d_D and d_N to classify prediction accuracy. The discriminators each contain three sub-discriminators that are 32x32 Patch GANs [15] with five convolution layers. The respective inputs to each sub-discriminator are a RGB image blurred with a 5x5 Gaussian kernel, a grayscale image containing only the luminance channel, and an image containing horizontal and vertical gradients [16].

The stereo networks for daytime and nighttime weights f_D,f_N based on the PSMNet [12] CNN contain three convolution layers, four residual blocks, and a Spatial Pyramid Pooling [17] module. The extracted features are concatenated into a 4D cost volume and then regularized by a 3D convolution architecture. Lastly, the features are passed through a disparity regression [18] layer to obtain the final disparity prediction.

Training Parameters: To train our model, we crop each image to 256×512. We train for 26 epochs and set the following parameters for our stereo GAN. Initial learning rate for the Adam [19] optimizer starts with $lr = 0.0002$ for the first 20 epochs of training and linearly decays to 0 in the last 20 epochs. It has momentum weights $\beta_1 = 0.5$, $\beta_2 = 0.999$ and the cycle-consistency [20] loss multiplier is $\lambda_{cyc} = 10.0$. The disparity reconstruction loss multiplier is $\lambda_{ste} = 0.05$. The batch size is 4 and each translation is buffered with the 50 previous pairs to mitigate spikes in the model's loss function. Maximum disparity is set to 48 as far objects are not as important in autonomous driving.

4. RESULTS

We perform an experiment using real nighttime data from the Oxford RobotCar Dataset [6][7] with data covering a variety of scenarios including glare, motion blur, noise, and overexposed lights. After preprocessing the night images, we train for 26 epochs which takes 5 days to train on one Nvidia GeoForce 1080Ti GPU. Then, we run test images on the trained models using the process shown in Fig. 7 where g_N^*, g_D^*, and f_N^* are the optimized weights.

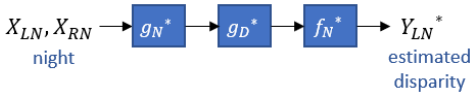
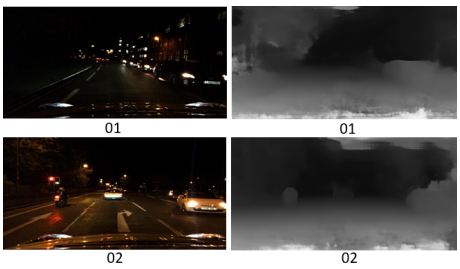


Fig. 7: Disparity Estimate from Nighttime Images

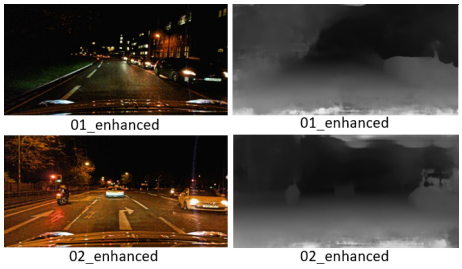
First, we trained a model with regular nighttime images for 40 epochs that achieves similar results to that in Sharma [10]. A sample of the resulting disparity map predictions for regular and enhanced night images are shown in Fig. 8 a-b. Visually, the regular images have more detailed edges around objects compared to the model trained on enhanced images.

Next, we trained a model with enhanced nighttime images for 26 epochs. The disparity map predictions of the same regular and enhanced nighttime images as before are passed through the model. The results are shown in Fig. 8 c-d.

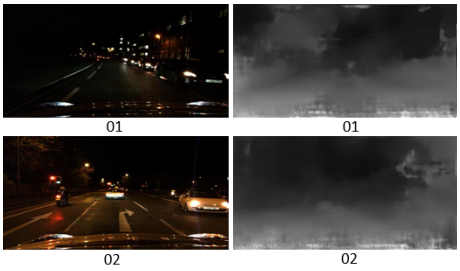
Error Evaluation: We report disparity error as the pixel percentage in the image that is k away from the true disparity. We vary k from 0 to 20, where $k = 0$ indicates correct disparity estimate and $k = 20$ indicates large error. Since the ground truth is sparse, we only test non-zero ground truth pixels. The bottom 10% of the disparity map is removed as it is covered by the car's hood and gives noisy results which are not present in the ground truth. Furthermore, the disparity estimates are



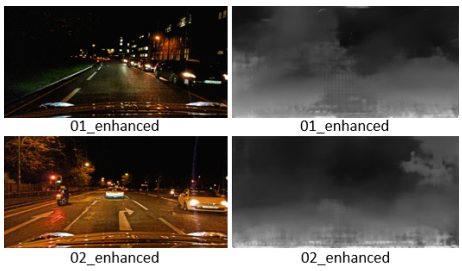
(a) Regular Night Image, Disparity Map



(b) Enhanced Night Image, Disparity Map



(c) Regular Night Image, Disparity Map



(d) Enhanced Night Image, Disparity Map

Fig. 8: Results for Model Trained with Regular Images (a-b) and Enhanced Images (c-d)

resized using nearest neighbor to match the ground truth size of 512×256. These steps are shown in Fig. 9.

A good disparity estimation algorithm yields a lower percentage error and higher percentage accuracy. We use an accumulated pixel percentage metric. Let $P(k)$ be the percentage pixel accuracy for a given k, then the accumulation function is defined as

$$Y(k) = Y(k-1) + P(k)$$

which yields $Y(0) = P(0), Y(1) = P(0) + P(1), Y(2) = P(0)+P(1)+P(2)$, and so on. The accumulated sum of pixel accuracy is shown in Fig. 10. The accuracy for regular and enhanced night images begin at 91% and 93% respectively and increases quickly to k=3 before gradually converging to 100%. This is because there are no incorrectly classified pixels at higher k. Enhanced images maintain higher accuracy compared to regular images up to k=5. Overall, the enhanced nighttime images enable the algorithm to obtain higher disparity estimation accuracy and classify more pixels correctly.

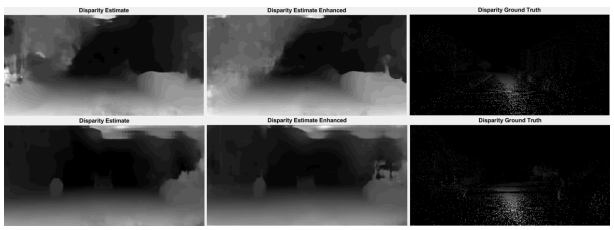


Fig. 9: Processed Disparity for Error Evaluation

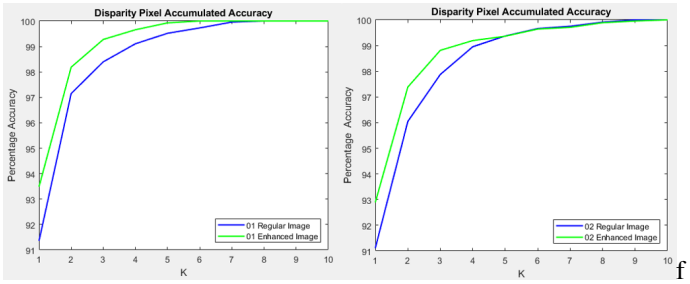


Fig. 10: Accuracy of Model Trained on Enhanced Images

5. CONCLUSION

In this paper, we proposed a disparity estimation algorithm for nighttime images. The nighttime image is first enhanced using a defogging and turbulence mitigation algorithm and then inputted to a generative adversarial network for training. Through experiments and error evaluation, the proposed network outperforms state-of-the-art algorithm for low-light depth estimation in a range of situations including glare, dark objects, noise, glow, and overexposed light. In addition to having less error, the resulting depth contains minimal noise.

This design improves the performance of current nighttime disparity estimation, but low-light depth estimation is still a largely untouched field. Further research towards low-light applications include object detection, semantic segmentation, or 3D reconstruction to improve safety for autonomous vehicles particularly when encountering pedestrians, vehicles, and guardrails.

References

[1] K. M. L. Y. S. Heo and S. U. Lee, "Joint depth map and color consistency estimation for stereo images with different illuminations and cameras," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013.

[2] L. H. Aashish Sharma, Loong-Fah Cheong and R. T. Tan, "Nighttime stereo depth estimation using joint translation-stereo learning: Light effects and uninformative regions," in *International Conference on 3D Vision (3DV)*, 2020.

[3] G. J. B. Clément Godard, Oisin Mac Aodha, "Unsupervised monocular depth estimation with left-right consistency," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[4] G. C. I. R. Ravi Garg, Vijay Kumar BG, "Unsupervised cnn for single view depth estimation: Geometry to the rescue," in *European Conference on Computer Vision (ECCV)*, 2016.

[5] D. G. B. H. M. C. K. Q. W. Yan Wang, Wei-Lun Chao, "Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[6] C. L. P. N. W. Maddern, G. Pascoe, "1 year, 1000km: The oxford robotcar dataset," in *International Journal of Robotics Research (IJRR)*, 2016.

[7] P. M. P. N. I. P. Dan Barnes, Matthew Gadd, "The oxford radar robotcar dataset: A radar extension to the oxford robotcar dataset," 2020.

[8] K. B. Gibson, "An analysis and method for contrast enhancement turbulence mitigation methods," 2013.

[9] M. M. B. X. D. W.-F. S. O. A. C. Y. B. I. Goodfellow, J. Pouget-Abadie, "Generative adversarial nets," in *Advances in Neural Information, Processing Systems*, 2014, p. 2672–2680.

[10] L.-F. C. Aashish Sharma, Robby T. Tan, "Depth estimation in nighttime using stereo-consistent cyclic translations," 2019.

[11] R. T.-M. P. L. V. G. Asha Anoosheh, Torsten Sattler, "Night-to-day image translation for retrieval-based localization," in *International Conference on Robotics and Automation (ICRA)*, 2019.

[12] Y.-S. C. J.-R. Chang, "Pyramid stereo matching network," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, p. 5410–5418.

[13] J. M. D. Chen Fu, Christoph Mertz, "Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[14] S. R. J. S. K. He, X. Zhang, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[15] T. Z. P. Isola, J.-Y. Zhu and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[16] R. T. M. P. L. V. G. A. Anoosheh, T. Sattler, "Night-to-day image translation for retrieval-based localization," 2018.

[17] S. R. J. S. K. He, X. Zhang, "Spatial pyramid pooling in deep convolutional networks for visual recognition," in *European Conference on Computer Vision (ECCV)*, 2014, p. 346–361.

[18] S. D. P. H. R. K. A. B. A. Kendall, H. Martirosyan and A. Bry., "End-to-end learning of geometry and context for deep stereo regression," in *Clinical Orthopaedics and Related Research (CoRR)*, 2017.

[19] J. B. Diederik P. Kingma, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations (ICLR)*, vol. 5, 2014, p. 346–361.

[20] P. I. J.-Y. Zhu, T. Park and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *International Conference on Computer Vision (ICCV)*, 2017.