Monocular Thermal Camera Depth Estimation with Optical Flow for Autonomous Drones

Tomoyasu Shimada[1,†], Hiroki Nishikawa[2,†], Xiangbo Kong[1,*,†] and Hiroyuki Tomiyama[1,†]

[1]*Graduate School of Science and Engineering, Ritsumeikan University*

[2]*Graduate School of Information Science and Technology, Osaka University*

Abstract

In this work, we propose a depth estimation from a monocular thermal image for drones. The highest flight speed of a drone is generally approximate 22.2 m/s, and long-distant depth information is crucial for autonomous since if the long-distance information is unavailable, the drone flying at high speeds is prone to collision. However, the sensors which can measure long-distance are too heavy to be equipped on drones. Therefore, autonomous drones apply a depth estimation method from a monocular camera. However, autonomous drones using the usual monocular camera depth estimation method do not operate properly during nighttime flights. Therefore, we propose a depth estimation method using a thermal camera, which is capable of capturing details even at night. Depth estimation based on thermal images alone has the problem of accuracy degradation due to noise in the thermal images. Therefore, we propose a method to improve accuracy by embedding optical flow, focusing on the fact that the drone is moving.

Keywords

thermal camera depth estimation, AirSim, optical flow, Pix2Pix

1. Introduction

Drones have been popular in recent years since drones are expected to play many roles. Autonomous drones are being used to maximize the convenience of drones. In terms of autonomous flights of drones, collision avoidance has been indispensable and regarded as one of the crucial issues. Typically, conventional solutions have employed distance sensors [1, 2, 3, 4, 5, 6]. However, such sensors with high performance are usually too heavy, and power-consuming to equip on a drone. In contrast, low-performance depth sensors can hardly have long-distance vision with high accuracy and would rather increase the risk of collisions with objects. Therefore, these problems are solved by using a lightweight monocular camera to estimate long distances [7, 8, 9, 10]. However, these methods do not work well for nighttime estimation since they are designed for daytime. Therefore, the use of a thermal camera, which is capable of capturing details even at night, enables depth estimation compared to our previous

The 4th International Symposium on Advanced Technologies and Applications in the Internet of Things (ATAIT 2022), August 24-26, 2022, Ibaraki, Japan

*Corresponding author.

†These authors contributed equally.

tomoyasu.shimada@tomiyama-lab.org (T. Shimada); nishikawa.hiroki@ist.osaka-u.ac.jp (H. Nishikawa); kong@fc.ritsumei.ac.jp (X. Kong); ht@fc.ritsumei.ac.jp (H. Tomiyama)

methods [11]. However, since this method uses only the information from the thermal camera for depth estimation, it is affected by noise and other factors that reduce its accuracy.

Based on the above, this paper proposes long-range thermal depth estimation using optical flow for drones' safe night flight. In addition, the proposed method is higher accuracy and lower collision rate than a conventional method. Our contributions are the following three points.

- This paper is the first to propose a depth estimation method that embeds optical flow in thermal images.
- The proposed method using a thermal camera enables long-range estimation at night, which could not be achieved with monocular images.
- In addition, The proposed method improves the accuracy of conventional thermal image depth estimation, contributing to a reduction in drone collision rates.

The rest of this paper is organized as follows. Section 2 shows related work on autonomous drone methods and depth estimation. Section 3 describes a proposed method. Section 4 shows the experimental results and Section 6 concludes this paper.

2. Related Work

There has been a lot of work associated with autonomous drones for several decades. A great deal of work has been devoted to flight safety, which is especially needed to prevent collisions with objects. One of the collision avoidance methods is the vision-based method. Vision-based collision avoidance methods use LiDAR images, time-of-flight (ToF) images, stereo images, and monocular images. In [12, 1, 2], the authors propose collision avoidance methods for a drone by using LiDAR. However, mounting a high-performance sensor such as LiDAR on a drone would result in an increase in weight. Since drones fly on limited battery power, increased weight increases power consumption, which can lead to problems such as not being able to fly long distances or not being able to take off in the first place.

In order to tackle these issues, the approaches of [6, 4, 3] propose obstacle avoidance methods for a safe flight using a lightweight and small depth camera or a stereo camera. The presented methods make a drone possible to avoid obstacles on-the-fly by determining an optimum waypoint on depth images. The work in [4], which is inspired by [3, 6], proposes a collision avoidance algorithm. This algorithm divides the image from the depth camera into five horizontal sections and determines the furthest section of the image as the direction of the drones. However, depth cameras with relatively low weight such as Azure Kinect, which is measurable only up to 10 m, released by Microsoft can hardly be installed on a drone since the drone is forced to low-speed flight for safety [13]. In this context, depth estimation from a monocular camera, which can overlook farther than a depth camera, has been attractive.

In the studies [7, 8], the authors present Support Vector Machine (SVM) based on depth estimation methods. The systems estimate depth from manually created features each patch divided from a monocular image. In these methods, features are obtained by a pre-trained SVM classifier. However, the accuracy of these methods is low as a result of handheld training data. In [14, 15, 16, 17], the authors present based on Convolutional Neural Networks (CNN). The CNN-based methods improve the accuracy of depth estimation more than the SVM method in

[7, 8]. However, the accuracy is still not sufficient enough since the network of these methods in [14, 15, 16, 17] is too simple to estimate depth accurately. In [18], a transformer-based method improves accuracy over CNN-based depth estimation methods. The transformer-based method [18] is enough accurate for secure flight. However, the method has a so this method is seemingly not suitable for the system for small devices like drones. In the study [10], we propose a method to pre-process the input image by embedding optical flow information in the image using a lightweight CNN. Our previous method [10] improves accuracy using a model with a small computational load, Pix2Pix, and succeeds in reducing the drone collision rate. However, while this method works during the daytime, the inability to acquire object edges at night significantly reduces the accuracy of estimation in dark areas. Therefore, there is a need for a method to estimate depth from thermal images, which can depict the edges of objects even in dark places such as at night.

However, there is currently no publicly available dataset that pairs thermal and depth images. In the study [19], the authors propose a method for depth estimation from thermal images using self-supervised learning. However, it is concluded that this method can only estimate a distance of 15 m to 20 m so this method is not suitable for a drone flight. In our previous work [11], we created a dataset consisting of 8,000 pairs of thermal and depth images on AirSim [20] and showed that depth estimation is possible with Pix2Pix [21]. We also conclude that the model trained on the simulator can be used to estimate long-range depth with a real thermal camera. However, the accuracy of this work [11] is not sufficient. In this paper, inspired by our previous work [10], we propose a depth estimation method that improves accuracy and reduces collision rate by pre-processing the input images using AirSim dataset.

3. Monocular Depth Estimation with Optical Flow Using AirSim

In this section, we present the depth estimation method from a monocular thermal camera. We employ Pix2Pix [21], which is one of the image translation techniques, to generate a depth map since it is a lightweight network and accurate more than other lightweight networks. Figure 1 shows the system overview of our proposed method. As shown in Figure 1, the proposed method consists of three parts: In the first part, the system generates an optical flow map from two serial frames. Second, the system embeds the generated optical flow map into a monocular thermal image. Finally, the Pix2Pix-based model estimates a depth map from the optical flow embedded image. In the following, we detail each part of the proposed method.

3.1. AirSim Dataset

In this work, to evaluate flight performance, we employ AirSim [20] to create the dataset for training and testing. AirSim is the most visually and physically realistic drone simulator available. AirSim uses UE4 as its virtual environment and can obtain various information from each object's mesh information. For example, the images obtained by a drone on AirSim can include RGB images, depth information from object placement coordinates, semantic segmentation images from object IDs, and thermal images from temperature information.

Figure 2a shows a thermal image in the AirSim environment. Figure 2b shows a depth image in the AirSim environment. As shown in Figure 2, AirSim can obtain long-range depth and

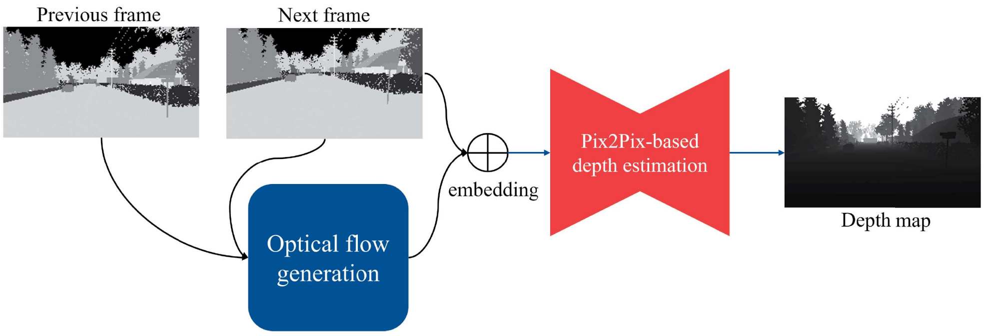

Figure 1: System overview of the proposed method [10].

Figure 2: AirSim [20] views:(**a**) Thermal image, (**b**) Depth map.

thermal images derived from mesh information at the same frame. In addition, since the drones in AirSim are assumed to be shooting from the same point, there is no need to account for parallax disparity due to sensor misalignment, which makes data set creation easy.

3.2. Optical Flow Generation and Embedding into Thermal Images

This method is inspired by the work [10]. We employ a general dense optical flow technique called Farnebäck method [22]. There is a method to generate sparse optical flow, but the sparse optical flow has lack information and cannot be effectively utilized.

We focused on the characteristics of the drone which is a mobile. In the drone view, close obstacles move larger and distant obstacles move smaller. Therefore, the optical flow displacement obtained from the drone viewpoint can be regarded as a simplified depth image. In the proposed method, a simplified depth image is created from the luminance values since the luminance values of the optical flow represent displacement. However, feature extraction is difficult to estimate depth from this simplified depth image because the simplified depth image cannot depict the edges of objects well.

In this work, we embed part of the pixel information of the simplified depth image into a thermal image to complement object edges and to make effective use of simple depth images. The concept of the proposed method is based on the atrous convolution in [23], which embeds the information of the simplified depth image into the thermal image. The simplified depth image is embedded at 5-pixel intervals so that the features of the original thermal image are

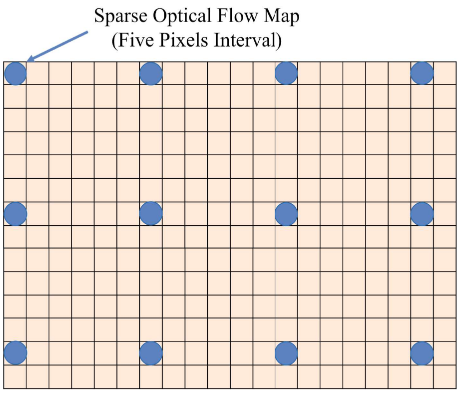

Figure 3: A concept of embedding a simplified depth image into a thermal image [10].

not lost. Figure 3 shows an example that a simplified depth image is embedded into a thermal image. Each pixel is embedded into the original thermal image. This embedding method is expressed in Equation (1).

$$E_{(n,m)} = \begin{cases} T_{(n,m)} & (n \bmod 5 \neq 0 \quad \cup \quad m \bmod 5 \neq 0 \quad \cup \quad S_{(n,m)} = 0) \\ S_{(n,m)} & (n \bmod 5 = 0 \quad \cap \quad m \bmod 5 = 0 \quad \cap \quad S_{(n,m)} \neq 0) \end{cases} \tag{1}$$

$E_{(n,m)}$ represents the pixel value of the simplified depth image at the pixel position of (n, m) embedded in the thermal image. $T_{(n,m)}$ is the pixel value of the thermal image at the (n, m) pixel position and $S_{(n,m)}$ is the pixel values of simplified depth image at the (n, m) pixel position. As shown in Equation (1), a simplified depth image is embedded in the thermal image at regular intervals, so the information in the simplified depth image without edges can be effectively utilized.

3.3. Pix2Pix-based Depth Estimation

In this work, our proposed method is based on Pix2Pix to generate a depth image from a thermal image [21]. Pix2Pix is a well-known image translation method based on CGAN [24]. CGAN consists of two networks such as a generator and a discriminator. The generator is used to generate images and The discriminator is used to identify real images and generated images. The generator of Pix2Pix consists of 7 convolution layers and 7 deconvolution layers for the encoder and decoder for a total of 14 layers. It also employs U-Net [25], which has skip connections between the encoder and the decoder at the same layer to prevent the lack of information during decoding.

The objective of the CGAN that we have employed is as shown in the following equation, which is referred to [21].

$$\begin{aligned} \mathscr{L}_{GAN}(G, D) = \mathbb{E}_{E,gt}[\log D(E, gt)] + \\ \mathbb{E}_{i,n}[\log(1 - D(E, G(E, n))] \end{aligned} \tag{2}$$

Here, E is an embedded image and Y is the ground truth. $D(E, Y)$ is the probability of identifying the ground truth as ground truth by the discriminator, and $D(E, G(E, n))$ is the probability of identifying the depth estimation image as ground truth by the discriminator. As shown in Equation (2), this loss function has the property that it gives large losses when the performance of the discriminator exceeds the performance of the generator, and small losses when the performance of the generator exceeds the performance of the discriminator. Therefore, by providing this loss to the Generator during training, it is possible to generate a plausible image. However, this loss depends on only two networks, so the accuracy of depth estimation does not improve much. In order to solve this problem, it is effective to add the following L1 norm to the objective of CGAN.

$$\mathcal{L}_{L1}(G) = \mathbb{E}_{E,Y,n}[\|Y - G(E,n)\|_1] \tag{3}$$

The L1 norm gives the absolute error between the ground-truth and depth-estimated images as a loss, allowing for absolute evaluation. The accuracy is expected to improve because the depth-estimated image works against the ground-truth error. By combining these two loss functions, the generator for depth estimation will be high accuracy. Therefore, the objective of Pix2Pix is as follows. w_{L1} is the weight of L1 norm. This parameter can be set during training.

$$G^* = arg \min_{\mathrm{G}} \max_{\mathrm{D}} \mathcal{L}_{CGAN}(G, D) + w_{L1}\mathcal{L}_{L1}(G) \tag{4}$$

The Discriminator learns to maximize this Equation (4), and the Generator learns to minimize it. Therefore, the Generator needs not only to make the Discriminator decide that the image is ground truth but also to generate an image that is closer to ground truth.

4. Experiments

In this section, we evaluate our method in terms of accuracy and performance to avoid collisions.

We use Intel Core i7-9700K (64 GB of main memory) and NVIDIA GeForce RTX 2070 SUPER. Dataset, which is used for training, validation, and testing, is collected from four maps provided in the AirSim environment; City Environment, Coastline, Neighborhood, and Soccer Field.

4.1. Comparison Accuracy between Proposed Method and Related Work

In order to evaluate the depth estimation error of models, we use rooted mean squared error (RMSE) and absolute relative error (Rel.) metrics. Equation (5) shows RMSE.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(v_i^Y - v_i)^2} \tag{5}$$

Here, v_i^Y is ground truth value, v_i is estimation value, and N is number of data. RMSE is used to measure how many outliers are in the depth estimation model. Equation (6) shows Rel..

$$Rel. = \frac{1}{N}\sum_{i=1}^{N}\frac{\|v_i^Y - v_i\|}{v_i^Y} \tag{6}$$

Specifically, the accuracy metrics are defined as:

$$\delta_n = \frac{\mathrm{Card}\left(\left\{y_i \,:\, max\left(\frac{y_i}{y_i^Y}, \frac{y_i^Y}{y_i}\right) < 1.25^n\right\}\right)}{\mathrm{Card}\left(\{y_i\}\right)} \qquad (n = 1, 2, 3) \tag{7}$$

This metric is called threshold accuracy. The threshold accuracy is higher is better. In this experiment, we will compare with the conventional method and evaluate the accuracy and error when the weights of L1 (w_{L1}) are changed. Table 1 shows the result of the evaluation of our proposed method and the comparison between the proposed method and a conventional method [11].

Table 1
Error and accuracy evaluation varying w_{L1} and other methods.

	w_{L1}	**Error (↓)**		**Accuracy (↑)**		
		RMSE	**Rel.**	δ_1	δ_2	δ_3
Pix2Pix[11]	100	7.083	0.536	0.702	0.820	0.893
proposed	0	7.118	0.548	0.669	0.783	0.867
	100	7.064	0.516	0.708	0.831	0.901
	200	6.986	0.528	0.710	0.832	0.901
	300	**6.906**	**0.418**	**0.720**	**0.832**	**0.908**

The error is the lower, the better, and the accuracy is the higher, the better. As shown in Table 1, when w_{L1} is zero, the error and the accuracy are the worst compared to other models. Then the method without optical flow [11] is better than w_{L1} is zero, and the method [11] is worse than the other models. The results show that increasing w_{L1} improves the accuracy and error.

First, consider that the accuracy and error were worst when w_{L1} was 0. The L1 norm calculates the absolute error and back-propagates to the model. On the other hand, the GAN loss is back propagated to the model using the results of the Discriminator, which quantifies plausibility, as the error. Therefore, it is thought that accuracy is improved by prioritizing absolute error over plausibility in depth estimation. Next, we discuss the superiority of the proposed method using optical flow over a conventional method [11]. Essentially, embedding a pixel from a simple depth image in a thermal image could result in an outlier in the accuracy of the corresponding pixel. However, CNN improves accuracy because it can take into account the pixel in question and surrounding pixels. Therefore, even if the pixel in question is an outlier, the accuracy of the surrounding pixels will be better than with conventional methods.

The following subsection describes the performance of these models when used on drones.

4.2. Safe Flight Evaluation in AirSim Environment

In this section, we perform a drone flight simulation using AirSim to demonstrate that the proposed method of depth estimation from thermal images can be a safe flight. In order to achieve the safe flight of an autonomous drone, it is necessary to plan a path for the drone to

Table 2
Comparison of collision rate.

Map	Collision Rate (%)	
	Pix2Pix[11]	Proposed
City environment	57.83	51.81
Coastline	51.25	31.75
Neighborhood	1.00	0.50
Soccer Field	5.50	5.25
AVERAGE	19.25	12.50

avoid collision with an obstacle itself, based on the depth estimation results. In the experiments, we use a state-of-the-art path planning method from depth map for autonomous drones, which is developed in [6]. We compare the collision rates represented by the number of collisions against the total number of flights. We define the collision rate as Equation (8)

$$Collision\ Rate = \frac{No.\ of\ Collisions}{No.\ of\ Flights\ (i.e.,\ 400\ flights\ in\ total)} \tag{8}$$

If the drone collides with an obstacle even once during a single flight, the number of collisions is counted as one. In the experiments, we use Windows 10Pro OS with 64GB main memory, Intel Core i7-9700K CPU, and NVIDIA GeForce RTX 2070 SUPER GPU. Depth estimation is processed by the CPU, as it is not consistent with the real drone when done on the GPU [10]. The flight simulation is conducted 400 times for each of the maps, which represent City, Coastline, Neighborhood, and Soccer Field. We compare two methods. The first method is the Pix2Pix-based method as a conventional method [11]. The second method is the proposed method that w_{L1} is 300. Table 2 shows the results of the collision rate in each map of AirSim.

As shown in Table 2, our proposed method makes it possible to obtain a lower collision rate than the method presented in [11]. Although the processing time is longer due to the optical flow processing, it is thought that the collision rate is reduced due to the highly improved accuracy.

5. Conclusion

We have developed an effective method for embedding optical flow diagrams into depth estimation using thermal images for dark environments such as nighttime. The collision rate of the proposed method achieves lower results than related works. Even using a lightweight model called Pix2Pix, we were able to improve the accuracy of thermal image depth estimation to a practically feasible level by devising input images. Even when Pix2Pix with the optical flow is used, the results show that there were few collisions. In order to implement the proposed method on actual drones, it is necessary to install a small high-performance computer such as the Jetson Xavier NX. Our future work is to study and develop a system of faster inference time so that it can be used in actual drones. The ablation study of our proposed method is also a future task. In addition, we will improve the network of depth estimation which can utilize simplified depth images and a thermal image in a real environment. Finally, we will experiment to evaluate the effectiveness of the proposed method in a real environment with actual drones.

Acknowledgments

This work is partly supported by KAKENHI 20K23333.

References

[1] Andrew Moffatt, Eric Platt, Brandon Mondragon, Aaron Kwok, Dennis Uryeu, and Subodh Bhandari, "Obstacle Detection and Avoidance System for Small UAVs Using A LiDAR," in *IEEE International Conference on Unmanned Aircraft Systems*, 2020.

[2] Yawei Hou, Zhenling Zhang, Chao Wang, Shouhu Cheng, and Demao Ye, "Research on Vehicle Identification Method and Vehicle Speed Measurement Method Based on Multi-rotor UAV Equipped with LiDAR," in *IEEE International Conference on Advanced Electronic Materials, Computers and Software Engineering*, 2020.

[3] Johann Borenstein and Yoram Koren, "The Vector Field Histogram-Fast Obstacle Avoidance for Mobile Robots," *IEEE Transactions on Robotics and Automation*, vol. 7, no. 3, pp. 278–288, 1991.

[4] Chenxiang Ma, You Zhou, and Zhiqiang Li, "A New Simulation Environment Based on AirSim, ROS, and PX4 for Quadcopter Aircrafts," in *International Conference on Control, Automation and Robotics*, 2020.

[5] Demetria Ma, Alexander Tran, Nick Keti, Ryan Yanagi, Peter Knight, Kedar Joglekar, Nicholas Tudor, Burt Cresta, and Subodh Bhandari, "Flight Test Validation of Collision Avoidance System for a Multicopter using Stereoscopic Vision," in *International Conference on Unmanned Aircraft Systems*, 2019.

[6] Erwin Perez, Alexander Winger, Alexander Tran, Carlos Garcia-Paredes, Niran Run, Nick Keti, Subodh Bhandari, and Amar Raheja, "Autonomous Collision Avoidance System for a Multicopter using Stereoscopic Vision," in *IEEE International Conference on Unmanned Aircraft Systems*, 2018.

[7] Kumar Bipin, Vishakh Duggal, and K. Madhava Krishna, "Autonomous Navigation of Generic Monocular Quadcopter in Natural Environment," in *IEEE International Conference on Robotics and Automation*, 2015.

[8] Yu-Hsun Lin, Wen-Huang Cheng, Hsin Miao, Tsung-Hao Ku, and Yung-Huan Hsieh, "Single Image Depth Estimation from Image Descriptors," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2012.

[9] Amir Atapour-Abarghouei and Toby P. Breckon, "Monocular Segment-Wise Depth: Monocular Depth Estimation Based on a Semantic Segmentation Prior," in *IEEE International Conference on Image Processing*, 2019.

[10] Tomoyasu Shimada, Hiroki Nishikawa, Xiangbo Kong, and Hiroyuki Tomiyama, "Pix2Pix-Based Monocular Depth Estimation for Drones with Optical Flow on AirSim," *Sensors*, vol. 22, no. 6, 2022.

[11] Tomoyasu Shimada, Hiroki Nishikawa, Xiangbo Kong, and Hiroyuki Tomiyama, "Depth

Estimation from Monocular Infrared Images for Autonomous Flight of Drones," in *International Conference on Electronics, Information, and Communication*, 2022.

[12] Joshua Redding, Jayesh Amin, Jovan Boskovic, Yeonsik Kang, Karl Hedrick, Jason Howlett, and Scott Poll, "A Real-Time Obstacle Detection and Reactive Path Planning System for Autonomous Small-Scale Helicopters," in *AIAA Guidance, Navigation and Control Conference and Exhibit*, 2007.

[13] Lan Anh Trinh, Nguyen Duc Thang, Duy Hau Nguyen Vu, and Tran Cong Hung, "Position Rectification with Depth Camera to Improve Odometry-based Localization," in *IEEE International Conference on Communications, Management and Telecommunications*, 2015.

[14] Shaoyong Zhang, Na Li, Chenchen Qiu, Zhibin Yu, Haiyong Zheng, and Bing Zheng, "Depth Map Prediction from A Single Image with Generative Adversarial Nets," *Multimedia Tools and Applications*, vol. 79, no. 21, pp. 14357–14374, 2020.

[15] F. Liu, C. Shen, G. Lin, and I. Reid, "Learning Depth from Single Monocular Images Using Deep Convolutional Neural Fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 10, 2016.

[16] M. Mancini, G. Costante, P. Valigi, and T. A. Ciarfuglia, "J-MOD2: Joint Monocular Obstacle Detection and Depth Estimation," *IEEE Robotics and Automation Letters*, vol. 3, no. 3, pp. 1490–1497, 2018.

[17] Kyle Hatch, John Mern, and Mykel Kochenderfer, "Obstacle Avoidance Using a Monocular Camera," *arXiv preprint arXiv:2012.01608*, 2020.

[18] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka, "Adabins: Depth Estimation Using Adaptive Bins," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2021.

[19] Xu Li, Meng Ding, Donghui Wei, Xiaozhou Wu, and Yunfeng Cao, "Estimate Depth Information from Monocular Infrared Images Based on Deep Learning," in *International Conference on Progress in Informatics and Computing*, 2020.

[20] Shital Shah, Debadeepta Dey, Chris Lovett, and Ashish Kapoor, "AirSim: High-Fidelity Visual and Physical Simulation for Autonomous Vehicles," in *Field and Service Robotics*, 2017.

[21] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros, "Image-to-Image Translation with Conditional Adversarial Networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[22] Gunnar Farnebäck, "Two-frame Motion Estimation Based on Polynomial Expansion," in *Scandinavian conference on Image analysis*, 2003.

[23] Jiayao Chen, Mingliang Zhou, Dongming Zhang, Hongwei Huang, and Fengshou Zhang, "Quantification of Water Inflow in Rock Tunnel Faces via Convolutional Neural Network Approach," *Automation in Construction*, vol. 123, pp. 103526, 2021.

[24] Mehdi Mirza and Simon Osindero, "Conditional Generative Adversarial Nets," *arXiv preprint arXiv:1411.1784*, 2014.

[25] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional Networks for Biomedical Image Segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, 2015.