2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
Macau, China, November 4-8, 2019

A Generative Model of Underwater Images for Active Landmark Detection and Docking

Shuang Liu, Mete Ozay, *Member, IEEE,* Hongli Xu, Yang Lin and Takayuki Okatani, *Member, IEEE*

***Abstract*— Underwater active landmarks (UALs) are widely used for short-range underwater navigation in underwater robotics tasks. Detection of UALs is challenging due to large variance of underwater illumination, water quality and change of camera viewpoint. Moreover, improvement of detection accuracy relies upon statistical diversity of images used to train detection models. We propose a generative adversarial network, called Tank-to-field GAN (T2FGAN), to learn generative models of underwater images, and use the learned models for data augmentation to improve detection accuracy. To this end, first a T2FGAN is trained using images of UALs captured in a tank. Then, the learned model of the T2FGAN is used to generate images of UALs according to different water quality, illumination, pose and landmark configurations (WIPCs). In experimental analyses, we first explore statistical properties of images of UALs generated by T2FGAN under various WIPCs for active landmark detection. Then, we use the generated images for training detection algorithms. Experimental results show that training detection algorithms using the generated images can improve detection accuracy. In field experiments, underwater docking tasks are successfully performed in a lake by employing detection models trained on datasets generated by T2FGAN.**

I. Introduction

Underwater landmarks are widely used for underwater navigation of autonomous underwater vehicles (AUVs) with short-range precision in many underwater robotics tasks, such as underwater docking [1]–[4], manipulation [5] and navigation [6]. Active landmarks [1], [6] such as light beacons, are preferred to passive landmarks [7]–[9] in underwater robotics tasks due to its relatively high visibility in water.

Robust detection of underwater landmarks is crucial for these tasks. Recently, remarkable progress has been made in general objection detection tasks using deep learning methods [10]–[12]. However, robust detection of active underwater landmarks is still challenging due to large variance in underwater illumination, camera viewpoint and water quality. Underwater illumination differs crucially at different time periods, such as daytime and night. Variance in viewpoints gives rise to deformation of shape of UALs. Water quality can vary according to water body, resulting in remarkable difference in imaging of landmarks. Underwater images captured in lakes tend to be greenish while blueish in sea water, since chlorophyll in phytoplankton absorbs strong blue and red wavelength, and concentration of chlorophyll in lakes is 10000 times more than those in ocean. All above pose great challenges to robust detection of UALs in various water body. An effective approach to solve this problem is to collect more diverse data. However, collecting data in wild underwater fields, e.g. lakes and ocean, is more time consuming and costly compared to collecting data in tanks, pools or air. Moreover, data acquired in fields may lack diversity in illumination and water quality.

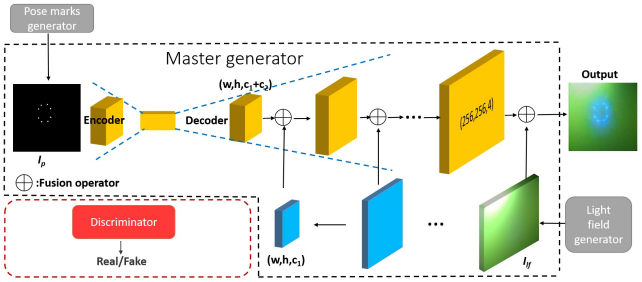

Fig. 1: An illustration of the architecture of the proposed Tank-to-field GAN (T2FGAN).

If a landmark detection model trained on data collected in a tank is used by AUVs for detection of UALs in wild environments, then their detection performance may decrease significantly, due to large gap between statistics of images captured in fields and tanks. A promising approach to alleviate this problem is to leverage power of synthetic data. Several works have utilized simulator engines by generating synthetic data. In [13], numerous vehicle images were generated according to various lighting and pose to facilitate learning of features more useful for detection of vehicles. Precarious pedestrian images, which are hard to obtain in real world datasets, were generated to improve accuracy of algorithms in [14]. Generative Adversarial Networks (GANs) [15] have produced impressive results in generating synthetic data. In [16], they improved vehicle detection performance in different domains using a network to transform images from a source domain to a target domain. Various GANs were proposed to generate images of human with different pose [17], [18].

In this work, we propose a network called Tank-to-field GAN (T2FGAN) (see Fig. 1) to improve accuracy of

Shuang Liu (corresponding author), Hongli Xu and Yang Lin are with State Key Laboratory of Robotics, Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang, China; Institutes for Robotics and Intelligent Manufacturing, Chinese Academy of Sciences, Shenyang, China; University of Chinese Academy of Sciences, Beijing, China e-mail: (liushuang1, xhl, liny@sia.cn).

Mete Ozay (corresponding author) is with Graduate School of Information Sciences, Tohoku University, Sendai, Miyagi, Japan email:(meteozay@gmail.com)

Takayuki Okatani is with Graduate School of Information Sciences, Tohoku University, Sendai, Miyagi, Japan; RIKEN Center for Advanced Intelligence Project, Tokyo, Japan email:(okatani@vision.is.tohoku.ac.jp)

detection models trained using images of UALs captured in tanks, and tested using those captured in wild fields. Our contribution is three-fold: 1) Our proposed T2FGAN is able to generate images of UALs according to different water quality, illumination, pose and landmark configurations (WIPCs), 2) T2FGANs can be used to generate ground truth of pose between cameras and landmarks, which is not possible to obtain in water, 3) T2FGAN can be used to increase statistical variance of training datasets considering statistical and physical properties of data collected in fields. Therefore, T2FGAN can help to improve detection accuracy of algorithms trained using these datasets in field experiments. These results are explored and demonstrated both by experiments on real world datasets obtained in the field, and underwater docking tasks performed in field experiments. The code is publicly available*.

II. Tank-to-field GANs

We propose a novel conditional GAN called Tank-to-field GAN (T2FGAN), which can generate images of UALs using various WIPCs. Thereby, we can bridge the gap between accuracy of detection methods trained using images of UALs captured in tanks, and tested using images captured in fields. Fig. 1 illustrates the overall architecture of the proposed T2FGAN, which contains four modules; a) a master generator (MG), b) a discriminator (D), c) a pose mark generator (PG) and d) a light field generator (LG). An MG is conditioned on a pose mark image and a light field image which is provided by a PG and a LG, respectively.

A. Master Generator, Fusion Operator and Discriminator

We assume that an underwater image I is composed of a foreground component I_f and a light field component I_{lf} weighted by a mask I_m. A master generator $MG : (I_p, I_{lf}) \mapsto I_r$ is used to learn a map from a pose mark image I_p and a light field image I_{lf} to an image of UALs I_r, by employing

$$I = I_f \odot (\mathbf{1} - I_m) + I_{lf} \odot I_m, \tag{1}$$

where $\odot$ denotes pixelwise multiplication, and $\tilde{I}_m$ is the complement of I_m. Two streams are devised for a master generator (MG) as illustrated in Fig. 1. The first stream, denoted by S_1, is responsible for generating I_f and I_m, conditioned on I_p provided by a PG. S_1 is an encoder-decoder based network with skip connections [19]. The encoder of S_1 firstly downsamples the input I_p to a bottleneck vector to learn efficient feature representations. Then, the bottleneck vector is decoded to I_f and I_m by the decoder. The second stream S_2 encodes light field information. Its encoded feature maps are fused with the layer-specific feature maps of the decoder in S_1 by our proposed fusion operator, which is explained next.

A fusion operator is proposed to implement (1) of the MG (see Fig. 2). A fusion operator takes feature maps of the decoder in S_1 denoted by $f_{s_1} \in \mathbb{R}^{w \times h \times (c_1+c_2)}$, and feature maps of S_2 denoted by $f_{s_2} \in \mathbb{R}^{w \times h \times c_1}$, as inputs. The first map f_{s_1} has the same width and height as the second map f_{s_2}, and the number of channels of f_{s_1} is twice that of f_{s_2}, since $c_1 = c_2$. The fusion operator decomposes f_{s_1} into two segments (subtensors). The first segment f_{c_1} contains the first half of channels of f_{s_1}. The second segment f_{c_2} contains the last half of channels of f_{s_1}. Then, f_{s_2} is integrated with these segments to generate a fusion map o by

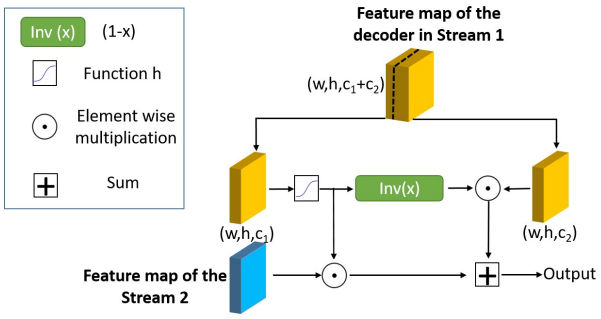

Fig. 2: An illustration of the proposed fusion operator.

$$o = h(f_{c_1}) \odot f_{s_2} + f_{c_2} \odot (\mathbf{1} - h(f_{c_1})), \tag{2}$$

where h is an activation function such as the sigmoid function.

We use $c_1 = 1$ and $c_2 = 3$ rather than $c_1 = c_2$ in the last fusion operator to enable T2FGAN learn the final foreground component I_f and the final fusion map I_m. Feature maps m_{last} are obtained after the last convolution layer in the decoder of S_1. The maps m_{last} are further integrated with the corresponding feature maps obtained from stream S_2 by (2), and the activation function h in (2) is computed by

$$h(x) = \frac{x - \min(x)}{\max(x) - \min(x)}, \tag{3}$$

where x is a feature map and $min(x)$ or $max(x)$ denotes the minimum or maximum value over all elements in x.

A discriminator D is used to discriminate real images and images generated by a generator. We used the same discriminator proposed in [20]. The light field image I_{lf}, pose mark image I_p, generated image $\hat{I}$ and real image I_r are concatenated as inputs to the discriminator in the training phase. The proposed generator MG and discriminator D are trained alternately by

$$MG^* = \min_{MG} \max_{D} \mathcal{L}_{cGAN}(MG, D) + \lambda \mathcal{L}_1(MG, I_r), \tag{4}$$

where $\mathcal{L}_{cGAN}(MG, D)$ is a loss function for the conditional GAN given in [20], and $\mathcal{L}_1$ is the ℓ_1 loss penalizing the difference between $MG(I_p, I_{lf})$ and the real image I_r. The parameter λ balances their contribution to the total loss.

B. Pose Mark Generator (PG)

PG provides pose mark images I_p to MG, which combines intrinsic properties of a camera, and the pose between the camera and UALs. In the training phase, I_p is obtained by manual labeling. In the inference phase, it is obtained by

*https://github.com/vincent341/T2FGAN

mapping 3D coordinates of UALs $\mathbf{P}_w$ to pose mark images I_p through the pinhole camera model

$$Z\mathbf{P}_{uv} = \mathbf{K}(\mathbf{R}\mathbf{P}_w + \mathbf{T}), \tag{5}$$

where Z is a scale factor, $\mathbf{P}_{uv}$ and $\mathbf{P}_w$ denotes coordinate vectors of the image and reference frame respectively, $\mathbf{K}$ is the intrinsic matrix of the camera, and $\mathbf{R}$ and $\mathbf{T}$ are rotation and translation matrices applied in the camera frame and reference frame, respectively.

C. Light Field Generator (LG)

Light field generator (LG) encodes illumination and water properties into the light field image I_{lf}. LG is implemented differently in training and inference phase. In the training phase, we consider estimation of light field information from an image of UALs as an image inpainting problem. We provide each underwater active landmark image with a manually labeled bounding box. Then, we utilize the inpainting algorithm proposed in [21] to generate patterns in the bounding box. An LG used in the training phase is denoted by LG_t.In the inference phase, LG maps the set of parameters θ to a light field image by $LG_{in} : \theta \mapsto l_{lf}$. We adopted an underwater image formation model proposed in [22] to compute LG_{in}. We denote a set of parameters by $\theta = \{E_l, P_l, \beta_r, \beta_g, \beta_b, g\}$, where E_l and P_l denotes the intensity and position of the light source respectively. g is the parameter of the Henyey-Greenstein phase function [23] which is used to model underwater scattering, and the parameter controls relative amount of forward and backward scattering. β_r, β_g, and β_b are extinction coefficients calculated for R, G, and B channels.

D. Estimating Illumination and Water Properties

We aim to bridge performance gap of detection methods trained and tested using images of UALs captured in tanks and fields. For this purpose, we propose to train detection methods using artificially generated images which have similar properties of images captured in fields.

To this end, we first estimate θ_i for the i^{th} field image $I_{f,i}$ by optimizing the objective function

$$\arg\min_{\theta_i} \|LG_{in}(\theta_i) - I_{f,i}\|_2. \tag{6}$$

θ_i is viewed as a sample of the random variable θ. θ_i is estimated from underwater images $I_{f,i}, \forall i$, containing only scattering components, obtained from Youtube. We optimize the objective function (6) using the Adam optimizer [24]. Next, we estimate distribution of θ. We assume that θ obeys a Normal distribution, $\theta \sim \mathcal{G}(\mu_\theta, \sigma_\theta^2)$. We estimated distribution of θ by fitting $\mathcal{G}(\mu_\theta, \sigma_\theta^2)$ to data. At last, we sampled θ from the estimated distribution $\mathcal{G}(\mu_\theta, \sigma_\theta^2)$ to obtain light field images in the target domain. We fed the sampled θ to LG_{in} to generate underwater light field images. Sample results obtained for a learned distribution of θ are given in Section III-C.

Fig. 3: *Left:* The experimental tank. *Middle and right:* Images of UALs with low and dense scattering, respectively.

III. Experimental Analyses and Results

A. Experimental Setup

Our aim is to first learn a generative model of images of UALs captured in a tank/pool. Next, we generate images of UALs under different WIPCs using the learned model. Then, we train a detection algorithm using the generated images in order to improve its detection accuracy in fields. To achieve this goal, we first collected images of UALs consisting of 8 blue led lights in a tank. The size of the tank is $0.85m \times 0.60m \times 0.45m$, as shown in Fig. 3. We covered all sides of the tank except the top side with opaque fabric to shield the tank from ambient light. We also added some flour to the water to increase scattering of the water body. Images of landmarks captured before and after adding flour are shown in Fig. 3. We captured 177 images of UALs with various poses in the tank. These images were used to construct a dataset denoted by D_{tank}.

B. Generating Images of UALs under Various WIPCs

Water Properties: Imaging conditions differ in various underwater environments due to variation of constituents of the water body of the environment. We parameterize water properties by $\theta_w = \{\beta_r, \beta_g, \beta_b, g\} \subset \theta$, whose definitions are given in Section II-C. β_r, β_g and β_b indicate how much light intensity is attenuated during propagation in R, G and B channels. Most sea water is blueish while lake water is greenish [25] as discussed in the introduction section. We generate blueish and greenish images of UALs by adopting extinction parameters $\beta_r, \beta_g, \beta_b$ given in [25], as shown in Fig. 4. g is a parameter of the phase function which controls the relative amount of forward and backward scattering. Fig. 4 shows landmark images generated with different g.

Illumination: Illumination in water is characterized by light source intensity E_l and light source position P_l. We assume that intensity is same in R, G and B channels. Fig. 5 and Fig. 6 shows UALs generated according to different light intensity and spatial configuration (position), respectively.

Pose: Pose refers to a rigid body transformation between a camera and landmarks, determined by rotation matrix R and translation matrix T given in (5). We employ Euler angles to represent rotation between the camera and the landmarks for ease of understanding in experiments. Fig. 7 shows sample images of UALs generated with various pose.

Configuration of Landmarks: Configuration of landmarks affects their detection and pose estimation accuracy. Various configurations are designed in previous works [3], [4], [26], [27]. In this section, we show how T2FGAN can

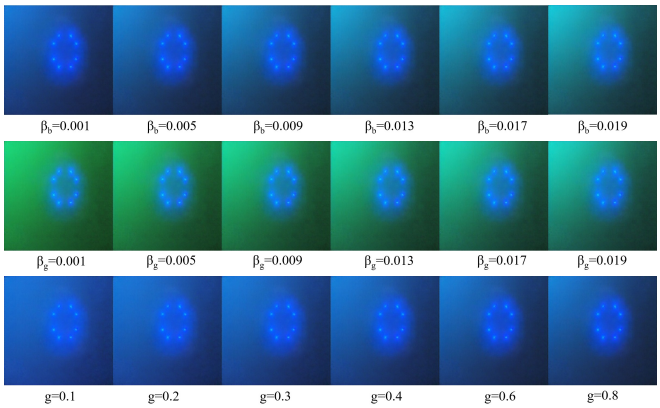

Fig. 4: Generated blueish ($\beta_r = 0.3, \beta_g = 0.026$, *the first row*) and greenish ($\beta_r = 0.3, \beta_b = 0.026$, *the second row*) images of UALs. *The last row:* Images of UALs generated with different g ($\beta_r = 0.3, \beta_g = 0.026, \beta_b = 0.003$).

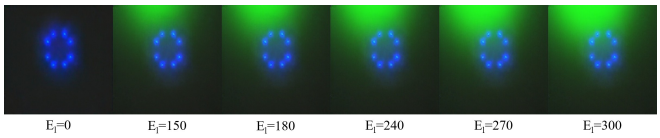

Fig. 5: Images of UALs generated according to different light intensity.

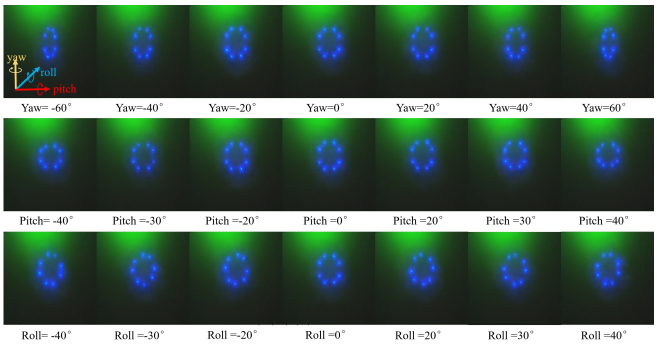

Fig. 7: Images of UALs generated with various poses. The first, second and third row shows results where only yaw, only pitch and only roll changes, respectively. The other Euler angles that are not depicted are set to 0. The coordinate is also attached to the first image.

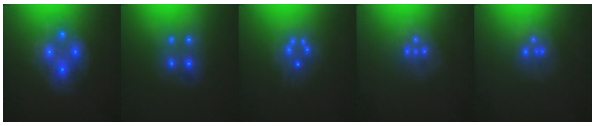

Fig. 8: Each figure from left to right shows a spatial configuration (position) of lights used in [4], [26], [3] and [27], respectively. The rightmost one is another view of the fourth figure.

generate images of UALs under various configurations. Our proposed method also enables other researchers to evaluate performance of algorithms with different configurations. Fig. 8 shows sample images of UALs generated using configurations suggested in previous works [4], [26], [3] and [27].

C. Improvement of Detection Accuracy

We analyze improvement of detection performance of SSD [28] with Mobilenet [29] trained on datasets D_{tank} and D_{gen} which were generated by a T2FGAN. We denote the detector trained on D_{tank} and D_{gen} by M_{tank} and M_{gen}, respectively. To construct D_{gen}, we first train the T2FGAN on D_{tank}. Then, distributions of parameters determining water properties $\theta_w = \{\beta_r, \beta_g, \beta_b, g\}$ in lakes are learned

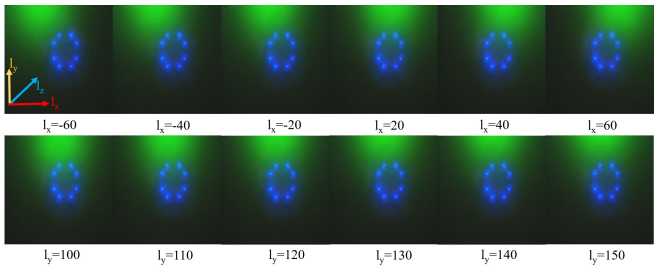

Fig. 6: Images of UALs generated according to different position of lights. The first and second row shows results where position changes along only l_x and only l_y, respectively. The coordinate is also attached to the first image.

from underwater videos captured in lakes obtained from Youtube[†‡] as described in Section II-D, since the test set is also collected in a lake. In total, 88 images that only contain the scattering component are selected from videos for learning distributions of θ_w. The learned distributions of θ_w are shown in Fig. 10. At last, T2FGAN generates synthetic images with particular WIPCs determined by θ_w drawn from the learned distributions. Other parameters belonging to the θ, except for θ_w, are drawn uniformly within a range. Then, 15699 images were generated for D_{gen}.

Detection performances of M_{tank} and M_{gen} were evaluated on the dataset D_{lake} using average precision (AP). D_{lake} consists of 19900 images of UALs, which were obtained by 8 blue led lights under various illumination and viewpoints. The dataset was collected in Qiandao lake, China, and used for underwater docking. Fig. 9 shows the precision-recall curve of M_{tank} and M_{gen}. It shows that the AP of M_{tank} is 93.61%, while 98.06% is obtained by M_{gen}, that is, AP was improved by 4.45%. This result suggests that T2FGAN can improve detection accuracy, and bridge the gap of performance of detection algorithms employed in tanks and fields, effectively.

We provide some typical results that are detected incorrectly by using M_{tank} but correctly by using M_{gen} in Fig. 11. Results show that M_{tank} tends to detect luminous objects. M_{tank} may work in underwater environments whose illumination is simple but degrades dramatically in

[†]https://youtube.com/watch?v=1hG0ciVKd94
[‡]https://www.youtube.com/watch?v=l_-DTkxPE3o

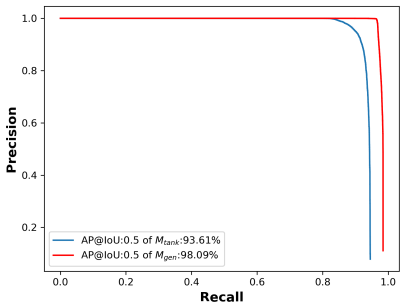

Fig. 9: Precision-recall of detectors M_{tank} and M_{gen}.

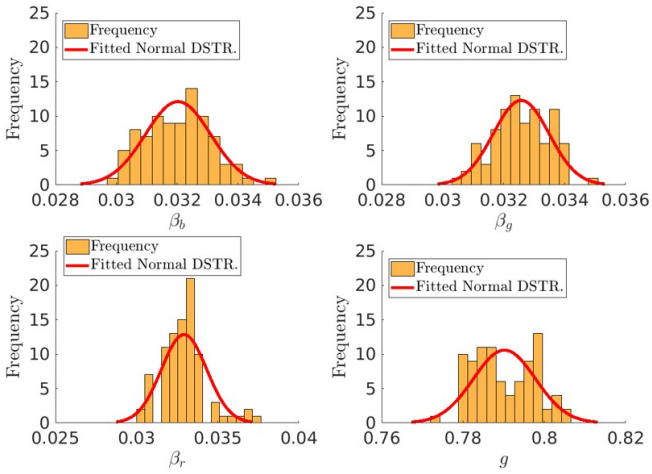

Fig. 10: Histograms and learned distributions of θ_w.

complex underwater environments, such as water body near the surface in daytime. Instead, M_{gen} learns more essential features of UALs from statistically diverse images of UALs generated by our proposed T2FGAN. We further validate the effectiveness of M_{gen} in an underwater docking task performed in real world field experiments in Section III-D.

D. Field Experiments

T2FGAN was further validated by a vision-based underwater docking (VBUD) task in field experiments, in Qiandao lake, China. The docking station and AUV in [30] were leveraged in field experiments. VBUD method consists of a detection and a pose estimation module. We leveraged M_{gen} as the detection module, and the pose estimation module proposed in [30] for pose estimation. The VBUD module was employed to perform final stage of underwater docking with short-range precision. An acoustic sensor, *Evologics S2C R*, was combined for rough long-range navigation.

Four consecutive runs of underwater docking of AUVs were performed, and all of them docked successfully. The trajectories of the AUV in the docking coordinates and the docking process of Run 1 are given in Fig. 12. The AUV started a docking process with random heading at random locations which were approximately $100m - 200m$ away from a docking station. It first leveraged positions offered by an acoustic sensor at a long distance, and switched to

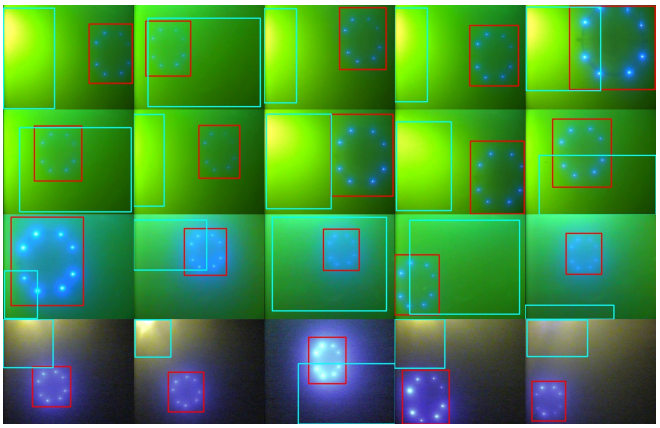

Fig. 11: Detection results provided by M_{tank} (blue boundingboxes) and M_{gen} (red boundingboxes).

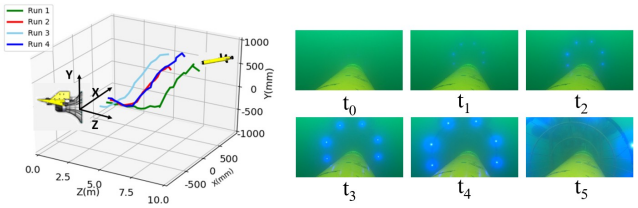

Fig. 12: *Left*: AUV trajectories obtained during vision-based docking in docking coordinates estimated by the vision-based algorithm. *Right*: Sample images (video frames) showing the docking process which were captured for the first run.

visual navigation once the docking station was detected by M_{gen}.

IV. Conclusion

In this work, a novel generative adversarial network called Tank-to-field GAN is proposed to generate images of UALs according to varying water quality, illumination, pose and landmark configurations of UALs to perform underwater robotics tasks.

In the experiments, we employed the proposed Tank-to-field GAN to generate images for data augmentation, pose estimation and training of detection algorithms. Experimental results show that we can improve baseline detection accuracy of a detection algorithm by training the algorithm using images generated by Tank-to-field GANs. In field experiments conducted in a lake, we performed four consecutive runs of underwater docking successfully using detection algorithms trained using the generated image datasets.

In current work, Tank-to-field GAN generates images of UALs using prior knowledge learned from streaming underwater images. In the future work, we plan to integrate the detection algorithm to Tank-to-field GAN by end-to-end training to improve accuracy of detection algorithms adversarially.

ACKNOWLEDGMENT

This work was partly supported by JST CREST Grant Number JPMJCR14D1.

REFERENCES

[1] P. W. Kimball, E. B. Clark, M. Scully, K. Richmond, C. Flesher, L. E. Lindzey, J. Harman, K. Huffstutler, J. Lawrence, S. Lelievre *et al.*, "The artemis under-ice auv docking system," *Journal of Field Robotics*, vol. 35, no. 2, pp. 299–308, 2018.

[2] Y.-H. Hong, J.-Y. Kim, J.-h. Oh, P.-M. Lee, B.-H. Jeon, K.-H. Oh *et al.*, "Development of the homing and docking algorithm for auv," in *The Thirteenth International Offshore and Polar Engineering Conference*. International Society of Offshore and Polar Engineers, 2003.

[3] J.-Y. Park, B.-h. Jun, P.-m. Lee, and J. Oh, "Experiments on vision guided docking of an autonomous underwater vehicle using one camera," *Ocean Engineering*, vol. 36, no. 1, pp. 48–61, 2009.

[4] Y. Li, Y. Jiang, J. Cao, B. Wang, and Y. Li, "Auv docking experiments based on vision positioning using two cameras," *Ocean Engineering*, vol. 110, pp. 163–173, 2015.

[5] N. Palomeras, A. Penalver, M. Massot-Campos, G. Vallicrosa, P. L. Negre, J. J. Fernández, P. Ridao, P. J. Sanz, G. Oliver-Codina, and A. Palomer, "I-auv docking and intervention in a subsea panel," in *Intelligent Robots and Systems (IROS 2014), 2014 IEEE/RSJ International Conference on*. IEEE, 2014, pp. 2279–2285.

[6] N. Gracias, J. Bosch, and M. E. Karim, "Pose estimation for underwater vehicles using light beacons," *IFAC-PapersOnLine*, vol. 48, no. 2, pp. 70–75, 2015.

[7] J. Sattar and G. Dudek, "On the performance of color tracking algorithms for underwater robots under varying lighting and visibility," in *Robotics and Automation, 2006. ICRA 2006. Proceedings 2006 IEEE International Conference on*. IEEE, 2006, pp. 3550–3555.

[8] F. D. Maire, D. Prasser, M. Dunbabin, and M. Dawson, "A vision based target detection system for docking of an autonomous underwater vehicle," in *Proceedings of the 2009 Australasion Conference on Robotics and Automation*. Australian Robotics and Automation Association, 2009.

[9] M. Wirtz, M. Hildebrandt, and C. Gaudig, "Design and test of a robust docking system for hovering auvs," in *Oceans, 2012*. IEEE, 2012, pp. 1–6.

[10] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *CoRR*, vol. abs/1804.02767, 2018.

[11] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. E. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *ECCV*, 2016.

[12] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, pp. 1137–1149, 2015.

[13] J. Tremblay, A. Prakash, D. Acuna, M. Brophy, V. Jampani, C. Anil, T. To, E. Cameracci, S. Boochoon, and S. Birchfield, "Training deep networks with synthetic data: Bridging the reality gap by domain randomization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 969–977.

[14] S. Huang and D. Ramanan, "Expecting the unexpected: Training detectors for unusual pedestrians with adversarial imposters," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, 2017.

[15] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[16] S.-W. Huang, C.-T. Lin, S.-P. Chen, Y.-Y. Wu, P.-H. Hsu, and S.-H. Lai, "Auggan: Cross domain adaptation with gan-based data augmentation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 718–731.

[17] L. Ma, X. Jia, Q. Sun, B. Schiele, T. Tuytelaars, and L. Van Gool, "Pose guided person image generation," in *Advances in Neural Information Processing Systems*, 2017, pp. 406–416.

[18] G. Balakrishnan, A. Zhao, A. V. Dalca, F. Durand, and J. V. Guttag, "Synthesizing images of humans in unseen poses," *CoRR*, vol. abs/1804.07739, 2018.

[19] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2015, pp. 234–241.

[20] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017, pp. 5967–5976.

[21] J. Ho Lee, I. Choi, and M. H. Kim, "Laplacian patch-based image synthesis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2727–2735.

[22] J. Tian, Z. Murez, T. Cui, Z. Zhang, D. J. Kriegman, and R. Ramamoorthi, "Depth and image restoration from light field in a scattering medium." in *ICCV*, 2017, pp. 2420–2429.

[23] L. G. Henyey and J. L. Greenstein, "Diffuse radiation in the galaxy," *The Astrophysical Journal*, vol. 93, pp. 70–83, 1941.

[24] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[25] C. D. Mobley, *Light and water: radiative transfer in natural waters*. Academic press, 1994.

[26] G. D. Watt, A. R. Roy, J. Currie, C. B. Gillis, J. Giesbrecht, G. J. Heard, M. Birsan, M. L. Seto, J. A. Carretero, R. Dubay *et al.*, "A concept for docking a uuv with a slowly moving submarine under waves," *IEEE Journal of Oceanic Engineering*, vol. 41, no. 2, pp. 471–498, 2016.

[27] T. Maki, R. Shiroku, Y. Sato, T. Matsuda, T. Sakamaki, and T. Ura, "Docking method for hovering type auvs by acoustic and visual positioning," in *Underwater Technology Symposium (UT), 2013 IEEE International*. IEEE, 2013, pp. 1–6.

[28] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *European conference on computer vision*. Springer, 2016, pp. 21–37.

[29] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *CoRR*, vol. abs/1704.04861, 2017.

[30] S. Liu, M. Ozay, T. Okatani, H. Xu, K. Sun, and Y. Lin, "Detection and pose estimation for short-range vision-based underwater docking," *IEEE Access*, vol. 7, pp. 2720–2749, 2019.