

Unlimited Resolution Image Generation with R2D2-GANs*

Marija Jegorova¹, Antti Ilari Karjalainen², Jose Vazquez², Timothy M. Hospedales¹

Abstract—In this paper we present a novel simulation technique for generating high quality images of any predefined resolution. This method can be used to synthesize sonar scans of size equivalent to those collected during a full-length mission, with across track resolutions of any chosen magnitude. In essence, our model extends Generative Adversarial Networks (GANs) based architecture into a conditional recursive setting, that facilitates the continuity of the generated images. The data produced is continuous, realistically-looking, and can also be generated at least two times faster than the real speed of acquisition for the sonars with higher resolutions, such as EdgeTech. The seabed topography can be fully controlled by the user. The visual assessment tests demonstrate that humans cannot distinguish the simulated images from real. Moreover, experimental results suggest that in the absence of real data the autonomous recognition systems can benefit greatly from training with the synthetic data, produced by the R2D2-GANs.

I. INTRODUCTION

Underwater optical visibility is often impaired due to the effect called marine snow, shown in Figure 1 (left), especially in cold seas. Because of that sonars are the primary source of the sensory information for autonomous vehicles operating underwater. The real-life underwater data collection is expensive, time-consuming, and impossible to carry out in certain locations.

However, the vast amounts of data are necessary for automating a number of the data-intensive applications, such as training of autonomous target recognition systems (ATR), as well as training human operators. The data shortage can be addressed by using the limited available data to train a high-quality simulator, capable of producing realistically looking synthetic imagery.

In this paper we propose a technique for synthesis of full-mission-long high-resolution seabed sonar scans, building upon the previously presented method, MC-pix2pix [1], suitable for lower resolution sonars. We suggest to extend the principle of MC-pix2pix further, enabling the resulting generative model not only to preserve all the strengths of its predecessor, but also to generate the data of any chosen high resolution, in principle - any desired resolution.

The speed of the data generation depends on the hardware, sonar range, and on the required resolution. For instance, for Marine Sonic sonars (512 pixels across track $\times 2$ channels) the generation rate is almost 20 times faster than the rate of the real data acquisition. For higher resolution sonars,

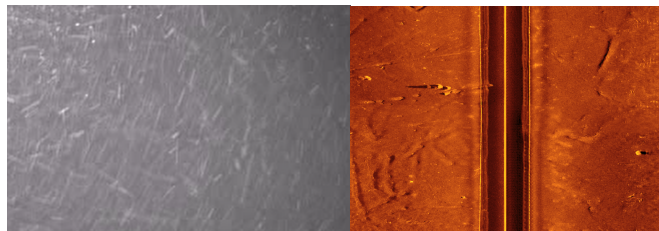


Fig. 1: **Examples of underwater sensors.** **Left:** optical camera - an example of visibility in northern seas, this effect is called *marine snow*, it is caused by high density of fine floats. In cases of poor visibility, sonars are often preferred over the cameras as more perceptually robust. **Right:** sonar side-scan. Port (left) is real, starboard (right) is synthesized.

like EdgeTech (approximately 4620 pixels across track $\times 2$ channels), the rate is at least twice faster than the rate of the real data acquisition. These estimates have been acquired with GTX 1080 Ti graphics card (12Gb RAM).

We call our method double-recursive double-discriminator Generative Adversarial Networks (R2D2-GANs, or “R2D2” for the sake of conciseness). To our knowledge, this is the first technique capable of adversarial generation of continuous and realistically-looking sonar side-scans of any requested size or resolution.

Potential applications of R2D2 can go far beyond the sonar imagery, as it can produce any type of large resolution imagery, provided a sufficient amount and quality of the initial training examples.

The visual examples of the results of the R2D2-GANs are provided in the Figure 2. Results demonstrated in this work are acquired with the image-to-image translation based architecture [2], which could be easily altered to accommodate another type of GAN, in order to better facilitate different simulation objectives.

II. RELATED WORK

The focus of this paper is the simulation of continuous side-scan sonar imagery of any requested size and resolution, with the user-controlled topography. Because of the visual nature and preferred stochasticity of such simulation, we focus on the corresponding family of the generative models.

Generative Adversarial Networks (GANs) were first introduced in 2014 [3]. Since then, they grew into a highly diverse class of methods and became the most popular way of the realistic image and video generation [4], image completion [5], super-resolution [6], and style transfer [7], [2], [8], [9]. There has also been a number of alternative applications, such as

* This work was supported by SeeByte Ltd

¹ University of Edinburgh, UK m.jegorova@ed.ac.uk, t.hospedales@ed.ac.uk

² Seebyte, UK antti.karjalainen@seebyte.com, jose.vazquez@seebyte.com

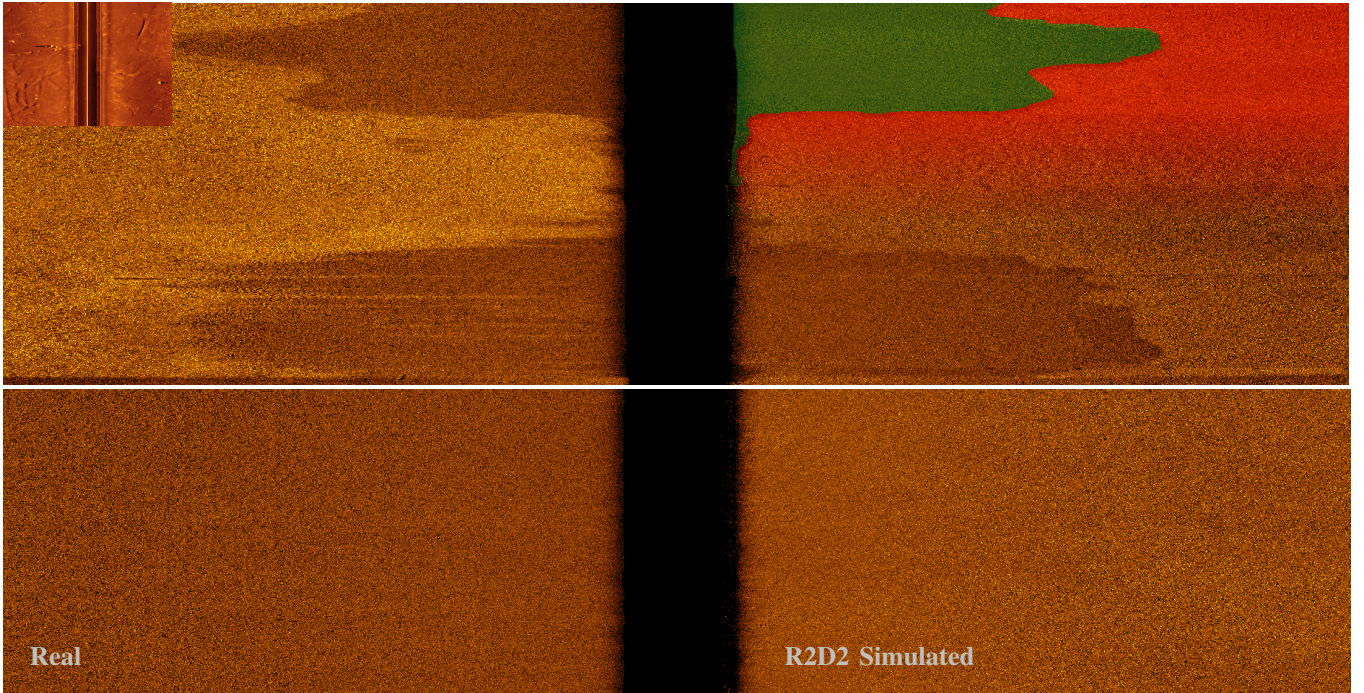


Fig. 2: **Visual results:** all images have the real sonar scans on the left, and the R2D2-simulated images on the right. The horizontal pairs of images correspond to the same semantic maps. The miniature image at the top-left corner is the Marine Sonic sonar data, generated with the MC-pix2pix method [1]. The rest of the images are EdgeTech sonar scans, generated with R2D2-GANs, provided in a relative scale according to the corresponding across track resolutions. The partial overlay in the top-right corner is an example of the semantic map, used by a generator network in order to control the topography of the simulation according to preferences of the user.

generation of socially acceptable trajectories [10] and control policy generation with GANs [11], [12]. Nevertheless, the visual data still stays the primary domain of application and development of the GAN models.

Despite that, there is still a relatively limited number of applications of GAN-based methods to the underwater sonar domain. Until recently the only application was the sonar imagery enhancement, mostly for the ATR training purposes - applying CycleGANs-powered style transfer to enhance the synthetic targets for the ATR training sets [13], and refining underwater video images [14]. However, there is almost no work focusing on the generation of the whole missions worth of the synthetic sonar data with complex terrains.

The first published work to bridge this exception was MC-pix2pix [1]. This method produces continuous full-mission-long sonar images for smaller across track resolution sonars (such as Marine Sonic). The results are both indistinguishable from real by human experts and capable of boosting the performance of the ATR systems. MC-pix2pix facilitates realistic conditional generation of the user-specified terrains with a modified pix2pix-style image translation. MC-pix2pix exploits Markov assumption for sequential generation of the image fragments in the along track direction, providing the continuity of the resulting image. An additional advantage of this piece-wise architecture is that it is relatively undemanding about the hardware. It is being supported even for the GPUs with very modest RAM capacities, which is almost never an option for the other higher-resolution GAN types.

Our new method, the R2D2-GAN, retains the general performance level of the MC-pix2pix method and completely surpasses the former on the magnitude of the across track image resolution it is capable of generating.

III. METHOD

The R2D2 belongs to the family of GANs. More specifically the R2D2 is an extension of the Markov-conditional image-to-image translation technique, MC-pix2pix [1], repurposed to enable image generation in larger resolutions.

GANs at a glance: the default GAN architecture usually consists of two neural networks. The first one, called the discriminator, learns to distinguish the real training images from the synthetic ones, whereas the second one, the generator, trains to create synthetic images that the discriminator cannot distinguish from real ones. Both networks are trained completely from scratch, gradually improving performance in an iterative manner via adversarial training. The final result of the GAN training is usually the trained generator network capable of generating diverse realistic images.

There are multiple conditional flavours of this basic architecture. Henceforth we focus on the paired image-to-image translation techniques, of which the pix2pix [2], [8] is a prime example. This choice is dictated by the requirement of the user-controlled topography.

The main features of the R2D2 technique: (i) incremental recursive generation (extending the Markov principle from [1]) applied along two axes (rather than just one, like in

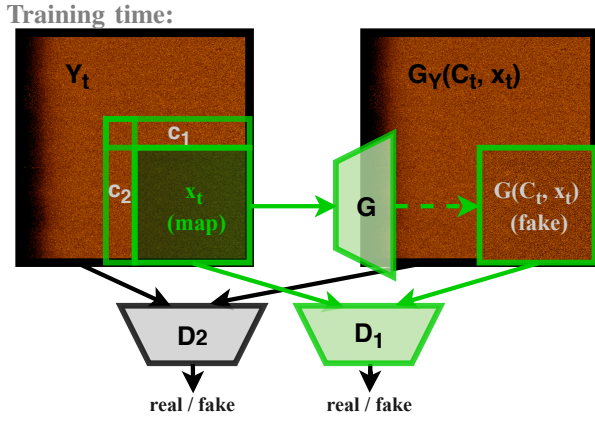


Fig. 3: **At training time:** inputs of the generator network include conditions c_1 and c_2 , which are small snippets of adjacent tiles (top and left) of the currently generated tile. The output of the generator G is the suggested synthetic image tile, generated taking into account the input conditions. The output of the generator is then assessed by the first discriminator D_1 along with real tiles. The second discriminator D_2 assesses the larger real image variants (2×2 tiles) - the unchanged and the edited with a generated tile. Both discriminators issue their decisions on whether the image is real or synthetic, and are rewarded based on the correctness of their decisions. Losses of both discriminators are then back-propagated through to the generator and used for the adversarial training.

[1]). This allows for handling any across track resolution. (ii) an additional discriminator is introduced for the coherence control of resulting larger scale images. Figures 3 and 4 provide the schematic illustration of the method.

At training time: as per scheme in Figure 3, first, the larger training images and their semantic maps are partitioned into multiple tiles. The generator inputs noise and the semantic maps of the current tile - for topography control. It also uses the additional conditions that include the location of a pixel in the across track direction and small snippets taken from the adjacent tiles above and to the left of the current tile. These are to ensure the continuity of the resulting large image. The generator is trained to produce realistic images given the above inputs and conditions.

The discriminator D_1 then tries to distinguish the real imagery from the simulated. The discriminator D_2 does the same, but processing the larger images (2×2 tiles) with the newly generated tiles embedded into them, and tries to distinguish them from the real unedited larger images.

The results provided in this paper are building upon the fully-convolutional pix2pix-style architecture with 9 resnet blocks [2], extended with additional conditions to support incremental recursive generation and additional “bigger picture” discriminator D_2 to further encourage the smoothness and continuity of the resulting image. This model is adversarially trained for 10 epochs with batch-size 3, and 3 gradient updates of discriminator D_1 corresponding to each gradient update of the generator. The training loss function can be

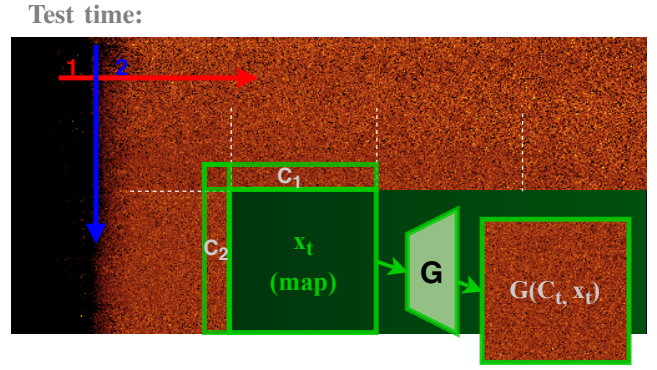


Fig. 4: **At test time:** only the generator is used at this stage. It produces image tiles first left-to-right, and then top-to-bottom. Each tile is conditioned on adjacent image snippets of the tile above and to the left of the currently generated tile. These conditions help to maintain the continuity of the larger picture produced at test time.

summarised as follows:

$$\begin{aligned}
 G_t^* = \arg \min_G \max_D \{ & \mathbb{E}_{C_t, x_t, y_t, z} [\|y_t - G(C_t, x_t, z)\|_1] \\
 + \frac{1}{2} (& \mathbb{E}_{C_t, x_t, y_t} [\log D_1(C_t, x_t, y_t)] + \mathbb{E}_{X_t, Y_t} [\log D_2(X_t, Y_t)] \\
 + \mathbb{E}_{z, C_t, x_t} [& 1 - \log D_1(C_t, x_t, G(C_t, x_t, z))] \\
 + \mathbb{E}_{C_t, z, X_t} [& 1 - \log D_2(X_t, G_{Y_t}(C_t, x_t, z))] \} \quad (1)
 \end{aligned}$$

where x_t are semantic maps and y_t are real sonar images per tile. X_t and Y_t are larger (2×2 tiles) semantic maps and sonar images respectively. X_t and Y_t include tiles x_t and y_t correspondingly. z is a random noise vector, and $C_t = [c_1, c_2]$ are the condition variables for the generator. $G_{Y_t}(C_t, x_t, z)$ stands for a larger image Y_t (2×2 tiles), where the native tile y_t is replaced by the generated tile $G(C_t, x_t, z)$. The first line of Equation (1) represents the L1 loss, a regularization term meant to reduce the blurring in the generator output [2]. The second line stands for the losses of both discriminators classifying the real data, and the last two lines are their losses of classifying the generated data.

At test time: the trained generator produces the entire image continuously piece by piece, first left-to-right, and then top-to-bottom, in accordance with the requested semantic maps. Refer to the Figure 4 for the schematic explanation.

Generalisation: the underlying technique of the R2D2 can be generalised beyond the specific GAN architecture. Nearly any GAN-based network, depending on the objectives and constraints of a specific task, can in principle be extended for the incremental recursive generation in a manner similar to the R2D2. The results provided in this paper are based on extending pix2pix-style architecture [2] solely for the purpose of providing the user with full control over the topography of the synthesized mission via utilisation of semantic maps.

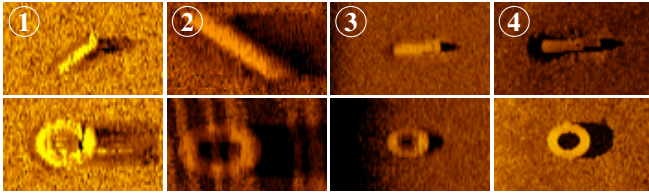


Fig. 5: **Examples of target objects (tyres and cylinders) and seabed types used for training the ATR:** 1. Uniform random noise background. 2. SonarSim-generated terrains¹. 3. R2D2-generated terrains. 4. Real terrains. All of these targets are inserted into the terrains using the Cycle-GAN-based technique [13], because of the limited availability of the real data with real targets.

IV. EXPERIMENTAL SETUP

Visual quality tests: a number of assessments were conducted in order to quantify the realism of the obtained imagery. We invited 10 domain experts to evaluate a selection of synthetic images created with the R2D2 and with classic pix2pix (for comparison), along with the real images. Participants inspected these images, labelling them as “real” or “synthetic” (“fake”).

All the image sets of the compared methods were presented in even proportions. Furthermore, all of the image sets (both real and synthetic) correspond to exactly the same set of the semantic maps to ensure the best possible comparability. Images acquired from the different sources were shown sequentially, one at a time, in order to avoid the cognitive bias. For the same reason there was no prior information provided on the proportion of real vs. synthetic images. The only information provided was that the test set contains both real and synthetic images.

Although the time taken to inspect each particular image was recorded and analysed, there was no time constraints imposed on the participants during the test time.

Autonomous target recognition (ATR) training: we argue that our proposed technique does not solely look good to human eye, but also can be of help for the autonomous systems training. For instance, assuming the lack of training data for the ATR, one could boost the training with the R2D2-generated data. Unfortunately, we do not possess unrestricted real data for sonars in EdgeTech resolution range (4620 or higher across track resolution), that would contain any real objects. Which is why we use the Cycle-GAN-based technique [13] to embed the artificial objects. In our specific case - cylinders and tyres, see Figure 5. In fact, there is a very small amount of unrestricted real seabed scans available, so we have to use what little real data we have for the test set.

There are a few training sets: (i) uniform random noise background, (ii) SonarSim¹ terrains - flat and rippled (respectively easier and harder for ATR to learn to operate on), and (iii) R2D2-generated terrains. Finally, we test the trained

¹ SonarSim - standard hard-coded and vaguely realistic side-scan simulator as used in [13], capable of generating various seabed textures with limited user control over the type of generated data, but not the exact topography.

Metrics:	fake labelled ‘real’	accuracy	av. time
pix2pix	0.14	0.88	4.79
R2D2 unnormalized	0.26	0.82	4.80
R2D2	0.78	0.56	5.23

TABLE I: **Visual test results:** the experiment was conducted with participation of 10 human experts possessing the daily experience of dealing with the sonar imagery. They were shown an equal number of images generated by different sources (both simulated and real) and asked to label them as “real” or “fake”. Images coming from different sources were shown one after the other in a random order to mitigate the cognitive bias.

R2D2 images were labelled “real” in 78% cases, which compares well with the benchmarks, as well as with real images labelled “real” (90%). Humans also were able to distinguish it from real with accuracy of 56 %, which is close to random chance in a two-class problem (“real” / “fake”). The last column of the results shows how long (in seconds) on average it took to classify an image. R2D2-produced images took significantly longer to process, compared to the other methods.

ATR performance on the small amount of the real sonar images we have available.

We use a simple ResNet-type architecture for ATR, trained from scratch for this experiment. Both the training and the test sets are rather small, due to the unavailability of the real data. Note that we do not claim the state-of-art level of ATR results here, only the relative benefit of using the R2D2-generated data in the absence of the real training data.

V. RESULTS

The visual results of the R2D2-GANs are shown in Figure 2, there is very little to no difference between the real (left) and synthetic (right) images. Also, please note the relative difference in scales between the typical data generated with MC-pix2pix (top-left corner) and R2D2-GANs (right). Because of the iterative recursive generation along both axes, R2D2 is practically unlimited in the data resolution it is able to generate. Naturally, there is always a trade-off between the magnitude of the generated image resolution and the generation speed.

The image assessment scores by human experts, presented in the Table I, are based on the results collected from 10 human experts with various level of expertise, but dealing with sonar imagery on a daily basis. The individual assessment metrics are as follows:

(i) R2D2 synthetic imagery is labelled “real” by humans in 78% of the cases. This is the highest score across the competing methods, which also compares reasonably well to 90% score for the real images classified as real.

(ii) Human classification accuracy being close to 50%, i.e. near random chance for two-class problem (“real” / “fake”), indicated inability of humans to tell apart the real and synthetic images. The R2D2 shows the lowest human classification accuracy score of 56%. This is comparable with

Train set:	Noise	SonarSim (flat)	SonarSim (rippled)	R2D2-GAN
Recall	0.00	0.3314	0.2255	0.4843
F1	0.00	0.4895	0.3653	0.6073

TABLE II: **Autonomous Target Recognition experimental results:** unfortunately, we have no access to unrestricted data with real targets at our disposal. However, we demonstrate, that expanding the ATR training datasets with R2D2-generated data may help the learning process. Potentially, a higher variety of the terrains available at training time should help the ATR system to generalise better. Fortunately, there is a Cycle-GAN-based method [13] to insert some artificial objects into the terrains, that is useful in this case. The training is conducted on 4 different types of terrains: uniform random noise, SonarSim¹ flat and rippled terrains (respectively less and more challenging for the ATR), and the R2D2-generated terrains. We train a simple ResNet-type network over these, and test on the real data with artificial targets embedded.

The results suggest that random noise in this case is completely useless, whereas R2D2-GAN on the contrary performs better than the competitors.

the 52% score obtained by the MC-pix2pix in an identical experiment [1], which is a remarkable result considering the significantly higher complexity of the current task of the higher-resolution generation.

(iii) We do not attribute any definite meaning to the average time spent on inspection of each separate image. Nevertheless, images produced by R2D2-GANs take the longest to classify. We suggest to interpret this as the R2D2-generated synthetic imagery posing the higher challenge for distinguishing it from real.

Our method outperforms the original pix2pix according to all the metrics in this assessment. It is also comparable with the current state of art - MC-pix2pix [1], which achieves the human labelling accuracy of 52% for images of much smaller resolutions. The R2D2, however, surpasses this competitor by the resulting resolution of the complete images it is capable of generating, while maintaining the comparable quality of the generated results.

ATR performance results are available in Table II. Both recall and F1-score² suggest that the R2D2-generated terrains provide significantly better training material compared to the random noise and SonarSim simulator.

VI. CONCLUSIONS

This paper presents the R2D2-GANs - a novel technique for generating the realistic synthetic imagery of any specified resolution and topography. This work provides both the quantitative and qualitative evidence confirming the realism of the images produced with our method. The empirical assessment also suggests significant advantages for the ATR systems trained with R2D2-generated data.

² F1-score - harmonic mean between the precision and the recall of the ATR system. Higher values correspond to the better performance.

The presented technique is in principle compatible with nearly any type of GANs, which might be of benefit for alternative objectives than those explored in this work. Thus providing the user with the ultimate control over the exact nature of the preferred data generation process. The R2D2-GANs are practically unlimited in the image resolution they can generate (at expense of the generation speed), which makes them immediately applicable to even higher resolution sonars, such as the Synthetic Aperture Sonars. Nonetheless, in the future work we intend to optimise this method further to make sure the fastest possible speed of the image generation for even higher resolutions.

VII. ACKNOWLEDGEMENTS

We would like to express our gratitude to Stephanos Loizou for his help with the ATR experiments, as well as the employees of the Seebyte Ltd who were of great help in gathering the data for the Table I by participating in the image assessment tests.

REFERENCES

- [1] M. Jegorova, A. I. Karjalainen, and J. Vazquez, "Full-scale continuous synthetic sonar data generation with Markov-Conditional Generative Adversarial Networks," *International Conference on Robotics and Automation (ICRA)*, 2020.
- [2] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5967–5976, 2016.
- [3] 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 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2014, pp. 2672–2680.
- [4] H. Cai, C. Bai, Y.-W. Tai, and C.-K. Tang, "Deep video generation, prediction and completion of human action sequences," in *The European Conference on Computer Vision (ECCV)*, 2018.
- [5] J. Yu, Z. L. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Generative image inpainting with contextual attention," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
- [6] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," in *CVPR*, 2017, pp. 105–114.
- [7] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," *International Conference on Computer Vision (ICCV)*, pp. 2242–2251, 2017.
- [8] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional GANs," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.
- [9] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu, "Semantic image synthesis with spatially-adaptive normalization," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.
- [10] A. Gupta, J. Johnson, L. Fei-Fei, S. Savarese, and A. Alahi, "Social GAN: socially acceptable trajectories with generative adversarial networks," *CVPR*, 2018.
- [11] Y. Li, J. Song, and S. Ermon, "Inferring the latent structure of human decision-making from raw visual inputs," *ArXiv*, 2017.
- [12] M. Jegorova, S. Doncieux, and T. M. Hospedales, "Generative adversarial policy networks for behavioural repertoire," in *IEEE 9th International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*, 2019, pp. 320–326.
- [13] A. I. Karjalainen, R. Mitchell, and J. Vazquez, "Training and validation of automatic target recognition systems using generative adversarial networks," *Sensor Signal Processing for Defence*, 2019.
- [14] C. Fabbri, M. J. Islam, and J. Sattar, "Enhancing Underwater Imagery using Generative Adversarial Networks," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2018.