Conditioned deep feature consistent variational autoencoder for simulating realistic sonar images

Jeygopi Panisilvam[1], Miguel Castillón[2], Nicholas Lawrance[3], and Roland Siegwart[3]

OCEANS 2022, Hampton Roads | 978-1-6654-6809-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/OCEANS47191.2022.9977288

***Abstract*—Multibeam imaging sonar is one of the primary sensors for underwater navigation with uncrewed underwater vehicles (UUVs) due to the robustness to turbidity and variable lighting conditions that limit the applicability of standard cameras. However, the operating principles and noise models of real sensors make imaging sonar challenging to accurately simulate, and acquiring real images experimentally is difficult and costly. This paper presents an approach for transforming a synthetically generated input image into the textural domain of real sonar images using a variational autoencoder (VAE) with a modified loss function. This allows us to generate realistic sonar images of simulated scenarios emulating the texture of real acoustic images. As a result, large datasets can be created from a relatively small amount of real images, which can be later used in many downstream applications, ranging from evaluating data association algorithms to deep learning. The method was evaluated using an isolated real and simulated dataset that trained a separate convolutional neural network (CNN) to discern between images in the sonar domain and simulated images. The VAE has several advantages over a compared Cycle Consistent Generative Adversarial Network (CycleGAN) approach, including more texturally accurate generated images, and allowing for more variation in generated images.**

***Index Terms*—Marine robotics, sonar, variational autoencoder (VAE), realistic acoustic image generation**

I. Introduction

Due to the attenuation of the most commonly-used electromagnetic waves in water, many underwater vehicles use acoustic (audio) systems for communications and sensing [1]. Sonar (sound navigation and ranging) is one of the primary navigation systems for UUVs. In particular, many vehicles use forward-facing multibeam imaging sonar as a navigation aid for navigating complex underwater terrain [2]. However, compared to depth sensors and cameras in above-water domains, imaging sonar has only seen limited use in autonomous navigation, and in many instances sonar images are interpreted by a human.

There has been limited progress in mapping and navigation from imaging sonar due to factors such as the high noise ratio and elevation ambiguity in sonar images. One significant limitation is the challenge of extracting repeatable image features for use in localization and mapping techniques like SLAM and structure from motion [3]–[5]. Learned feature detector and descriptor approaches [6] have been successfully applied to camera, depth, and multi-modality matching. These approaches offer considerable promise for imaging sonar. However, they are typically based on training via large amounts of simulated and real images. A major limitation for sonar is collecting large datasets, due to the challenging conditions required for deploying imaging sonar in representative domains.

One approach to address this issue would be to use simulated sonar images. Although the principal of imaging sonar is relatively simple, faithfully reproducing the artifacts that occur in real sonar images is complex. Factors such as beam cross-talk, interfering returns and sensor noise depend on many factors and vary between specific sensors.

The current approach is inspired by previous work [7] that has addressed this problem by using the established CycleGAN [8] approach to generate realistic sonar images. CycleGANs produce realistic results on simulated images, however this approach produces some limitations.

In our work, we propose a VAE. This approach offers multiple advantages to the CycleGAN approach. First, the proposed method improves textural quality by learning the representation of texture through the training of the first VAEs network. Further, the trained conditional VAE is able to generate many output images from one simulated image input, as it takes a sample from a latent space distribution based on the provided input image. This second advantage is significant as it produces a much larger amount of generated images compared to the CycleGAN which typically deterministically generates only a single output image based on one input image.

II. Related Work

A. Paired Image Transforms

Initial work performed in this area focused on using the pix2pix architecture in order to generate realistic looking sonar images [9]–[11]. Systems trained using this type of network showed good results, but require paired data of matching simulated inputs and ground truth outputs. For underwater autonomous navigation, ground truth data is often not available, and hence paired datasets are difficult to acquire in most cases. As a result, more general methods need to be used in order to create a mapping from a simulated image to a sonar image where no ground truth paired structural result is available.

B. CycleGAN

The current state of the art method for converting simulated sonar images into realistic sonar images relies on the Cycle-

This work was supported by the ThinkSwiss Asia Pacific Scholarship program and armasuisse S+T under project number № 043-32.

[1] Principal author. Student, The University of Melbourne, Victoria, Australia, jpanisilvam@student.unimelb.edu.au

[2] Underwater Robotics Lab, University of Girona, Catalonia, Spain, miguel.castillon@udg.edu

[3] Autonomous Systems Lab, ETH Zürich, Zurich, Switzerland, {lawrancn, rsiegwart}@ethz.ch.

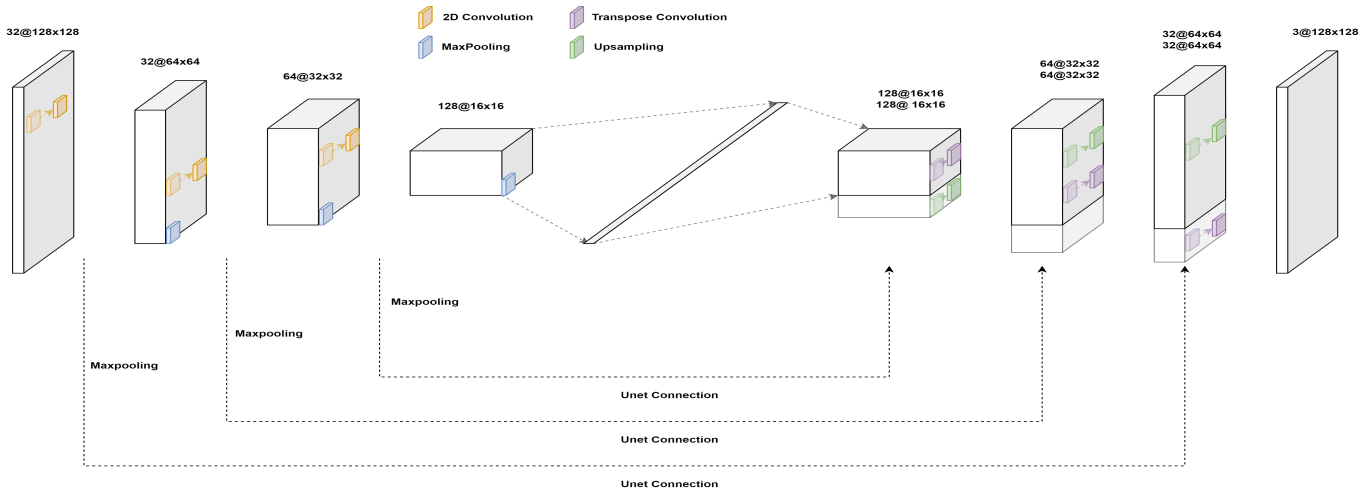

Fig. 1: The proposed VAE Network Architecture

GAN framework [12]. This approach attempts to reconstruct a real sonar image scene where the dimensions and distance of the objects in the scene are known quantities. Due to the availability of ground truth data, the paper is able to compare its approximated generated result to real sonar images, which it assumes to be a ground truth representation. Quantitative comparisons to ground truth data were evaluated in addition to qualitative assessments due to these factors. As CycleGANs produce a deterministic mapping between the image domains, it results in one generated image for each input image. The quantitative evaluation methods presented use several metrics including Mean Squared Error, Peak Signal to Noise Ratio, and Structural Similarity Index Measure [13].

III. Method

A. Dataset Acquisition

There are two datasets which are of importance in this research. The first dataset is the reference real sonar images which were captured using a 512 beam Blueprint Subsea Oculus 1200d imaging sonar [14]. The second dataset consists of the simulated images. These were created through the use of a pre-existing sonar simulator developed by the Tethys Lab using the Unity Game Engine. An algorithm was created to circle a set of objects and take persistent image captures of an artificial scene. In total there were 1732 simulated images used in training, and 2662 real sonar images used in training. A subset of this data was used for validation.

B. Conditioned Deep Feature Consistent VAE

The method presented leverages the modified loss metric from deep feature consistent VAEs [15] along with a unique training method to provide results with strong clarity and low diffusion (see Fig. 3). It is first proposed to generate an encoder-decoder VAEs scheme (where $D(\cdot)$ is the decoder, and $E(\cdot)$ is the encoder) such that the following condition is satisfied:

$$\min D(E(x)) - x, \text{ and } Z \sim \mathcal{N}(0,1) \tag{1}$$

$\forall x \in X$, where x is a random image sampled from the set of training images X, and Z is the representation of the latent space as a unit-variance normal distribution [16]. Following this, to generate image data in the natural sonar domain from simulated images, the simulated images are placed as a training input into the trained VAE (see Fig. 2). This results in the simulated image being encoded and decoded using the encoder and decoder weights learned from training sonar dataset. Simultaneously, the network is biased to learn the geometrical representations of the desired sonar objects. The perceptual loss function for both networks can be described as [15]:

$$L_p = \sum_i w_i L_p^i, \text{ with } L_p^i = ||\phi(x)^i - \phi(\hat{x})^i||^2 \tag{2}$$

where $\phi(x)^i$ and $\phi(\hat{x})^i$ are representations of the original image x and the reconstructed image $\hat{x}$ at each layer, with weight w_i.

The above perceptual loss is summed with the Kullback-Leibler (KL) divergence loss to ensure a matching distribution [17]. The KL divergence loss is given by:

$$L_{KL} = KL(q(t|x)||p(t)) \tag{3}$$

where $q(t|x)$ is the variational distribution and $p(t)$ is the prior over the latent variables. $q(t|x)$ is the learned distribution which is learned by the encoder network, and $p(t)$ is chosen to be the standard normal distribution.

The total loss of the system is given by:

$$L = L_p + L_{KL} \tag{4}$$

The latent dimension size is set to 500, with a learning rate of $5 \cdot 10^{-4}$. Dropout is used as a regularisation method. For both convolution and transpose convolutions, a leaky Rectified Linear Unit function is used as an activation layer. For feature extraction it was found that using the pretrained VGG19 model [18] excluding the output layer extracted features much quicker than training a feature extractor from scratch. As a result this is recommended during training.

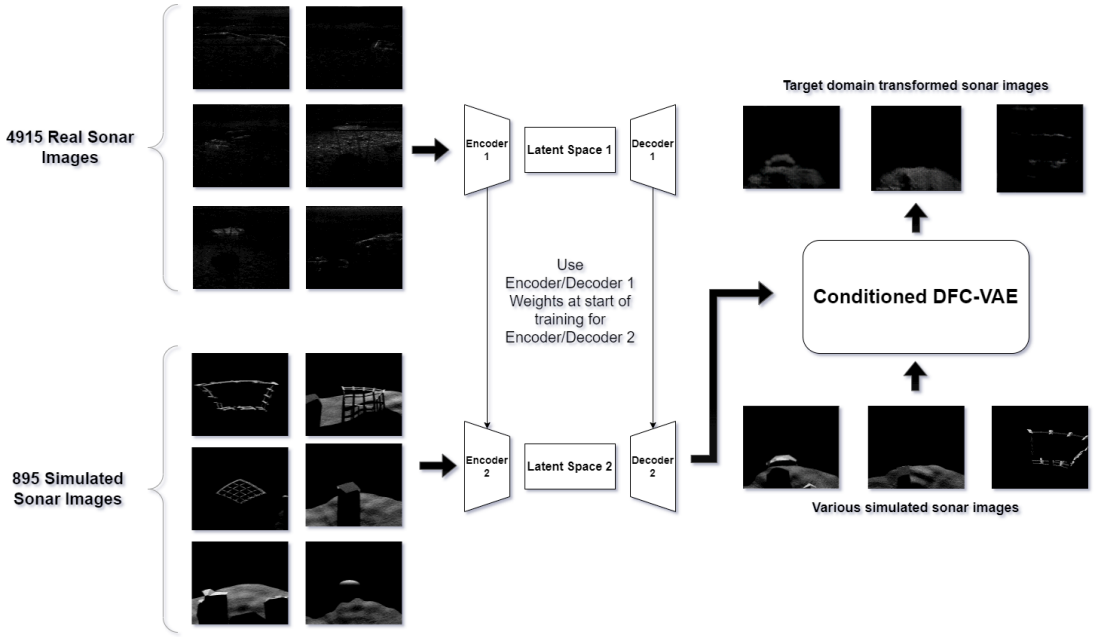

Fig. 2: Architecture and training pipeline of the proposed method.

Both training cycles use an early stopping mechanism to limit training networks after the requisite feature mappings have been learned. From experimentation, it was found that for the initial training cycle, 80 epochs was sufficient, and for the second training cycle, 60 epochs was sufficient. To draw random samples after the network was trained, a sample is drawn from the second input encoding, and is perturbed with some Gaussian noise variation given by $Z \sim \mathcal{N}(0, 1)$.

A significant advantage of the conditioned Deep Feature Consistent Variational Autoencoder (cDFCVAE) over the CycleGAN network structure is its resistance to mode collapse during training. Mode collapse occurs when a network learns to predict a very small subset of values that satisfy a specified loss metric. By "cheating" the task in this way, the network usually doesn't learn as much information as it should. The mode collapse issue is particularly prevalent in CycleGANs due to an imbalance in the strength of discriminators and generators learning representations of images. The cDFCVAE circumvents this issue by attempting to map the entire space of condensed images, and drawing a random sample from this space. This prevents mode collapse as long as the generated space is properly normalized. In addition to this, in the formulation of the VAE presented, early stopping is used to prevent the model from overfitting to either dataset. As a result losses do not become too low where the encoder or decoder is no longer able to learn any further information.

For the intended use case of this algorithm, it presents another significant advantage. As the generated images can be generated in a non-deterministic way through the use of random sampling, a large amount of generated images can be drawn from a single simulated input image. This is due to the fact that a one to many mapping is created between the latent space representation and the final transformed output.

IV. Results

The result of implementing the algorithm can be seen below (see Fig. 3,4). The results show excellent textural translation, and strong structural translation.

Evaluating nonstandard generated images is often challenging, especially when a ground truth result is not available. Initially, traditional generative image evaluation techniques such as the Frechet Inception Distance (FID), Inception score and KL Divergence were investigated. However on closer inspection, these methods were unsuitable for sonar images. This is because sonar image features are significantly different from standard images due to the method in which pictures are taken with sonar imaging equipment. Obtaining a FID score, Inception score or KL Divergence value would only provide information about how closely the generated sonar images match a set of organic real world images. As sonar images are not images taken with a standard imaging camera, these scores would not provide a fair comparison.

As an alternative, a CNN was constructed to check for accuracy by training the CNN on the unused segment of the dataset (see Fig. 5). The output neurons of the CNN were taken without a classification layer in order to determine how close to the space of real or fake images any given image was.

The modified optimisation problem used to determine the metrics present in Table I can be described by:

$$\max_{\bar{E},\bar{D}} \sum_{n} \mathcal{C}(\bar{E}(\bar{D}(y))) - \mathcal{C}(y) \tag{5}$$

where $\mathcal{C}$ is the CNN that the images are passed into to determine if they are a real or generated image.

The results in table I show that the cDFCVAE that has been designed performs better than a CycleGAN approach for the same problem, using the same dataset. A point to note in

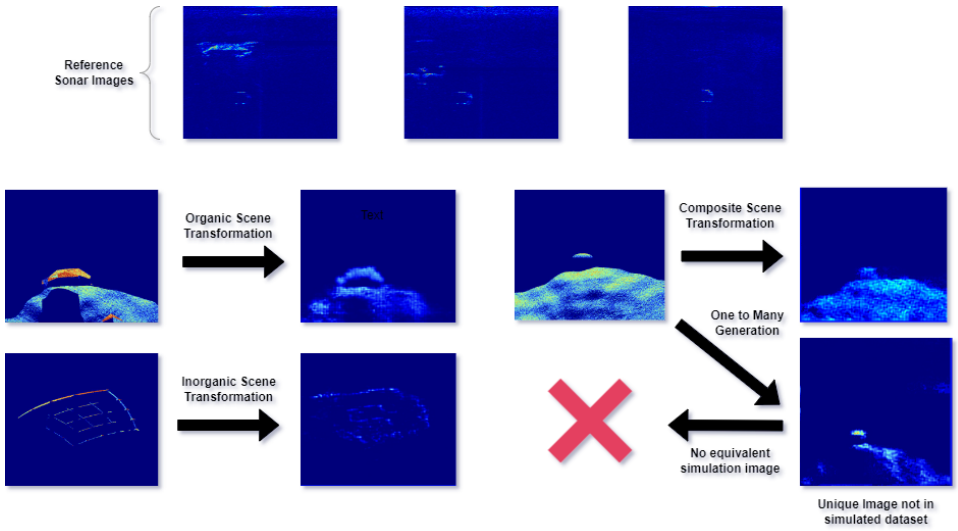

Fig. 3: On the left, organic and inorganic scene transformations. On the right, an example of the *one-to-many* result possible with our approach. Reference Sonar images are present at the top of the figure.

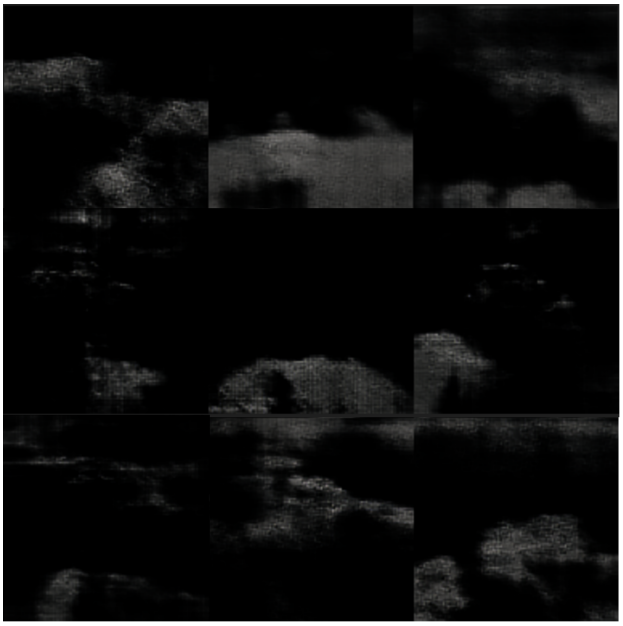

Fig. 4: Set of 9 different generated images. Several of the images have features that are not present in any of the datasets. This is a result of the sampling present from the VAE structure.

the evaluation method is the fact that the dataset for training images contained mainly constructed images (84%), and only a few organic images (16%). The performance on the network may be dependent on this factor, which is why there may be a significant difference between the accuracy metrics for constructed and organic images.

This evaluation method is still not sufficient to conclude that the method performs well, as the geometric structure of the real sonar images and simulated sonar images was often significantly different.

Due to the reasons above, resultant images were also qualitatively inspected, and it was noted that the results show

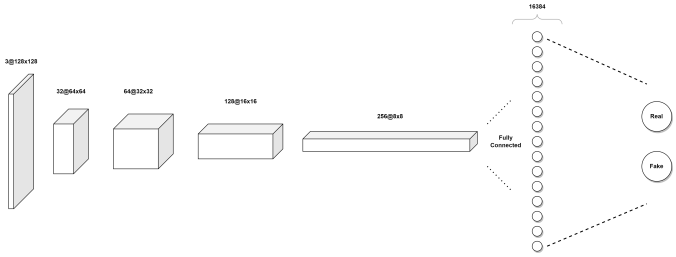

Fig. 5: CNN Structure used for classification

TABLE I: Comparison of CNN output between original and translated images. Outputs close to 1 indicate that the CNN classified most of the images as belonging to the dataset of real images.

	Raw Sim. Image	*CycleGAN*	*C-DFCVAE*
Organic Objects	0.114-0.212	0.442-0.843	0.507-0.803
Constructed Objects	0.188-0.281	0.266-0.652	0.552-0.867

excellent textural translation from the initial simulated sonar imagery. These results also appear consistent with the source image texture (see Fig. 3). Structural translation shows strong performance as well, with both major and minor features being translated across domains, and minimal feature loss.

V. Conclusion and Future Work

A method was presented to synthetically generate realistic looking sonar images from a relatively small amount of real images using a VAE. Our method replicates the textural quality of sonar images accurately, while also simultaneously allowing the creation of a large amount of generated images.

In the future, these results can be later used in many downstream applications, ranging from evaluating data association algorithms to deep learning. When used in this manner, the upscaling algorithm for the decoder will need to be evaluated. Many current upscaling algorithms suffer from 'checkerboard' artifacts usually observable in the Fourier domain of an image.

While these effects are subtle and difficult to observe with the human eye, algorithms will be able to pick up on the structured patterns caused by upscaling.

Another avenue for future research lies in the evaluation methods for the algorithm presented. As mentioned above, quantitative evaluation is one of the most challenging aspects of generative research, especially in the field of sonar imaging. If a strong quantitative evaluation metric could be determined to assess the quality of generation, the method could be cross evaluated and compared to other generative approaches. Being able to evaluate abstract generated structures would be a significant benefit to the further development of new computer vision algorithms for underwater autonomous navigation.

REFERENCES

[1] E. Gallimore, J. Partan, I. Vaughn, S. Singh, J. Shusta, and L. Freitag, "The WHOI micromodem-2: A scalable system for acoustic communications and networking," in *MTS/IEEE OCEANS*, 2010.

[2] C. Satheesh, S. Kamal, A. Mujeeb, and M. Supriya, "Passive sonar target classification using deep generative β-vae," *IEEE Signal Processing Letters*, vol. 28, pp. 808–812, 2021.

[3] M. D. Aykin and S. Negahdaripour, "On feature matching and image registration for two-dimensional forward-scan sonar imaging," *Journal of Field Robotics*, vol. 30, no. 4, pp. 602–623, 2013.

[4] T. A. Huang and M. Kaess, "Towards acoustic structure from motion for imaging sonar," in *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2015, pp. 758–765.

[5] J. Li, M. Kaess, R. M. Eustice, and M. Johnson-Roberson, "Pose-graph SLAM using forward-looking sonar," *IEEE Robotics and Automation Letters*, vol. 3, no. 3, pp. 2330–2337, 2018.

[6] D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superpoint: Self-supervised interest point detection and description," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2018, pp. 224–236.

[7] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *IEEE international conference on computer vision*, 2017, pp. 2223–2232.

[8] J. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," *CoRR*, vol. abs/1703.10593, 2017. [Online]. Available: http://arxiv.org/abs/1703.10593

[9] E.-H. lee, M. Jeon, H. Jang, B. Park, A. Kim, and S. Lee, "Study on the training effectiveness of deep learning with synthesized underwater sonar image using pix2pix and fcn," in *2020 IEEE/OES Autonomous Underwater Vehicles Symposium (AUV)*, 2020, pp. 1–3.

[10] M. Jegorova, A. I. Karjalainen, J. Vazquez, and T. Hospedales, "Full-scale continuous synthetic sonar data generation with markov conditional generative adversarial networks," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, 2020, pp. 3168–3174.

[11] L. Rixon Fuchs, C. Larsson, and A. Gällström, "Deep learning based technique for enhanced sonar imaging," *5th Underwater Acoustics Conference & Exhibition (UACE), Hersonissos, Crete, Greece*, vol. 1021-1028, 2019.

[12] D. Liu, Y. Wang, Y. Ji, H. Tsuchiya, A. Yamashita, and H. Asama, "Cyclegan-based realistic image dataset generation for forward-looking sonar," *Advanced Robotics*, vol. 35, no. 3-4, pp. 242–254, 2021.

[13] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[14] B. Schildknecht, "Feature-based estimation of 3D point clouds with forward-looking imaging sonar," 2020, BSc. thesis, ETH Zurich.

[15] X. Hou, L. Shen, K. Sun, and G. Qiu, "Deep feature consistent variational autoencoder," in *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2017, pp. 1133–1141.

[16] D. P. Kingma and M. Welling, "An introduction to variational autoencoders," *CoRR*, vol. abs/1906.02691, 2019. [Online]. Available: http://arxiv.org/abs/1906.02691

[17] A. Asperti and M. Trentin, "Balancing reconstruction error and kullback-leibler divergence in variational autoencoders," *IEEE Access*, vol. 8, pp. 199 440–199 448, 2020.

[18] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014. [Online]. Available: https://arxiv.org/abs/1409.1556