CNN-based Approach for Opti-Acoustic Reciprocal Feature Matching

Hyesu Jang[1], Giseop Kim[1], Yeongjun Lee[2], and Ayoung Kim[1*]

***Abstract*— Sonar and optical images are widely utilized imaging sensors in underwater research. Despite their popularity, the optic camera's limited view and acoustic camera's high level noise may hinder their utility in a single-sensor application. To alleviate such challenges, complementary exploitation of the two sensors was introduced and suggested for opti-acoustic approaches in underwater sensing. However, defining the features and matching the two different sensor modalities is challenging. In this paper, we propose to automatically match the features between the optic image and the acoustic image by using the opti-acoustic image conversion that is based on a neural network. The main focus of this paper is to generate the optic-style image, given an acoustic image via the Convolutional Neural Network (CNN)-based neural style transfer. In doing so, we acquired the contour of the object in sonar image with content feature extraction and an embodied optic image style with style feature extraction. The generated image has the content of the sonar image, yet, the style is the optic image. Finally, we verified the utility of out result by calculating the cosine similarity and the number of matched features between two images.**

I. Introduction

An optic camera and sonar are widely applied sensors that are used in underwater vision. Optic imaging is preferred in many terrestrial and underwater perception because the resulting images are the most intuitive vision sensors for a human. As can be found in the literature, most of the research in computer vision is developed with optic camera images. In underwater conditions, however, scattering light and turbid water disturb the camera when obtaining the desired images. With its acoustic-based sensing, sonar has advantages for capturing the underwater environment. Even under low-visibility water conditions, sonar is capable of detecting the target object from a longer range.

To complement the limited sensing modality from each sensor, researchers are constantly attempting to resolve the problem by investigating the correlation between clear data and environment. [1], [2], and [3] handled underwater image dehazing and restoration with optical camera, and [4] and [5] covered sonar image restoration and saliency extraction. We focused on multimodal sensor fusion problems, especially opti-acoustic stereo imaging, which was introduced by [6]. Interrelationship derivation between the optic and acoustic imaging sensors could resolve shortcomings in the sensing modality.

[1]Department of Civil and Environmental Engineering, Korea Advanced Institute of Science & Technology (KAIST), 291 Daehak-ro, Yuseong-gu, Daejeon 34141, Republic of Korea [iriter,paulgkim,ayoungk]@kaist.ac.kr

[2]Korea Research Institute Ship and Ocean engineering (KRISO), Daejeon, Korea 34103 [leeyoungjun@kriso.re.kr]

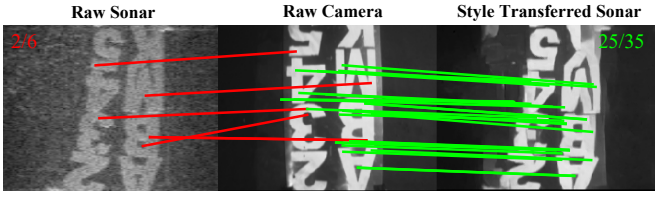


Fig. 1. Summary of CNN-based opti-acoustic image conversion. Before (left) and after (right) the transference, feature matching results have been substantially improved.

Since the first discovery of artificial neural networks, the research area of computer vision has become extremely broad. Neural style transfer is one area of computer vision research that has received compliments. As the right side of Fig. 1 illustrates, the neural style transfer reforms the content image to the given style by calculating two losses.

We propose a neural network based approach for the opti-acoustic feature matching problem. Both sensors have its style; the optic image is vivid and sonar is noisy. Even if the image captured the same target, the feature points from each image may be different. The objective of this paper is to find opti-acoustic correspondence using a neural-transferred image. We expect the neural style transferring to boost denoising and vivify sonar image. The following are presented and discussed in this paper.

- To prevent the noise-to-content conversion in the neural transferring process, we pre-processed the image so that the network concentrates only on the content object.
- Inter-transformation between sonar and the optic image with neural network training was conducted for opti-acoustic stereo imaging.
- We verified the opti-acoustic correspondence with cosine similarity and feature matching.

II. Related Works

A. Opti-acoustics

The opti-acoustic feature matching and image calibration was conducted in [6]. In their work, by utilizing the five known opti-acoustic correspondence in calibration step, they found the epipolar curve for the acoustic image. With the epipolar geometry and opti-acoustic correspondence, they attempted to perform 3D reconstruction. The opti-acoustic projection model based bundle adjustment was implemented in [7]. They applied various feature detectors/descriptors, such as Scale Invariant Feature Transform (SIFT), Binary Robust Invariant Scalable Keypoints (BRISK), Accelerated-KAZE (A-KAZE), and Dense Adaptive Self-Correlation (DASC) for opti-acoustic image matching. Consequentially, DASC presented the most satisfying opti-acoustic feature

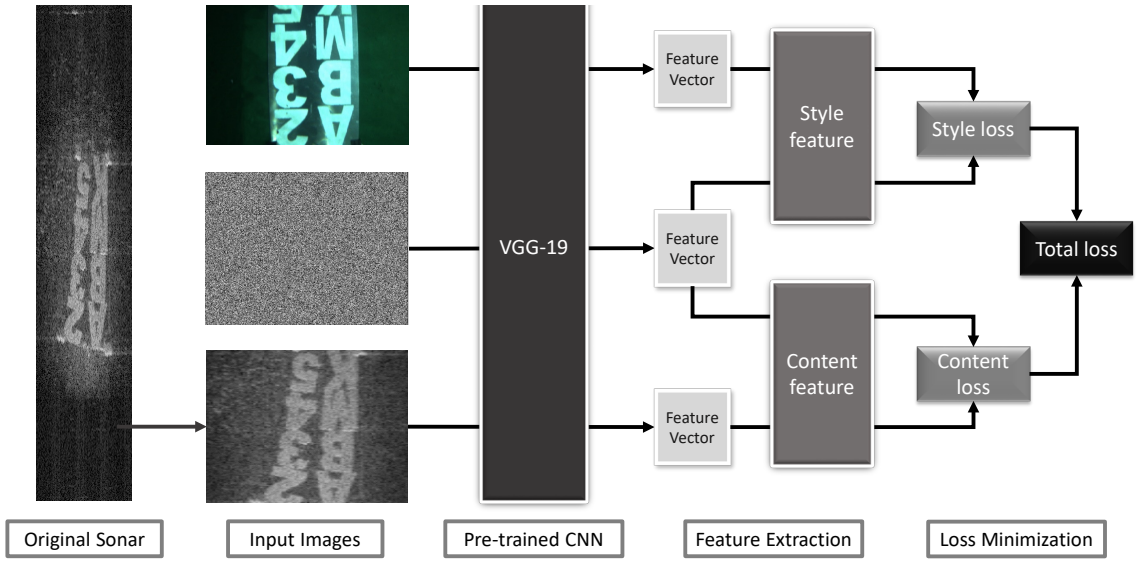


Fig. 2. Opti-acoustic style transfer algorithm. Original sonar image is processed to an input form. VGG-19 produces feature vectors. Style and content losses are calculated with feature maps. Finally network submit total loss for generated image.

matching result. However, these opti-acoustic studies assume that the optic camera and sonar are capturing the same scene. To apply existing computer vision research to the opti-acoustic data, automatic feature detection and matching have to be preceded. As stated in [7], the opti-acoustic feature matching problem is challenging due to the noise of the sonar image.

B. Neural Style Transfer

Neural style transfer is a recent arising study in both computer vision and robotics. While maintaining the image contents, the style of the image is modified with the characteristic of the input image. The existing neural style transfer method has two big branches; one is loss minimizing with pre-trained CNN, and the other is training the style transfer model with the Generative Adversarial Network (GAN).

In the work by [8], the authors extracted the style feature from the image with CNN and a Gram matrix. The VGG-19 from [9] was the key network, and the style of the image was decided with the comparison between the feature layers. Finally, the CNN-based neural style transfer algorithm was established in [10]. They merged the style and content features of the image with an appropriate rate and realized the style transfer of the image. The style transfer with GAN was introduced in [11] and [12], which is known as pix2pix and CycleGAN. They generated realistic images with training both a descriptor and a generator under the feedback system. They have generated decent image quality, however, the GAN-based style transfer requires intuitive image sets that represent the corresponding scene.

C. Application to Underwater Image

In underwater research, [3] proposed a GAN-based approach for underwater optic camera image restoration. They utilized a dataset with 6,500 optic images for network training and produced outstanding results. In the case of sonar, [5] scored the saliency of the sonar image with CNN to improve underwater simultaneous localization and mapping (SLAM) performance. As appear by other studies, previous network-based underwater research has focused on single-sensor processing. Therefore, this study focus on acoustic-to-optic image conversion with a neural network, which has not previously been attempted. To utilize GAN, preparing a sufficient image dataset is required in advance. However, this is challenging in underwater as the amount of public opti-acoustic image datasets is not enough to execute GAN at present. Thus, we opted to develop a CNN-based opti-acoustic style transfer.

III. Method

The proposed method is composed of image pre-processing, neural feature extraction, and image generation. In the image pre-processing phase, our intention is to prevent noise from becoming the main content of the image. Then, we conduct neural feature extraction with a pre-trained network. Each style and content feature can be derived from the layer of the pre-trained CNN network. Lastly, the image is generated with both style and content loss minimization. The learning process is divided into two steps, style and content loss definition, and total loss definition for back propagation. The overall process is depicted in Fig. 2.

A. Image Pre-processing

Fig. 3 describes a sample raw sonar image in which, the upper and lower parts depict highly noisy pixels without any meaningful features. In the first step, we eliminate meaningless regions in the sonar image to eschew noise-object transferring. Although automatic region detection is preferred for the multi-object case, we crop with respect to the center as only one target exists in our case. In this pre-processing phase, we crop the image by analyzing the characteristic of the sonar image.

B. Feature Extraction via CNN

Two methods are dominant in neural network based image style transfer. One is pre-trained CNN-based neural style transfer, and the other is style-reformed image generation with GAN. As pre-trained CNN transfer requires only two images for image generation, the required time for training is relatively short. However, GAN-based transfer demands a large amount of datasets, which affects the learning time. This need for large dataset may be a limitation for underwater application; no public sonar image dataset is readily available, unlike the wide accessibility of optic images (e.g., ImageNet or CIFAR-10). Furthermore, the most considerable advantage of GAN is an aesthetically refined result, which does not necessarily infer quantitative improvement in robot vision applications. Instead of GAN, therefore, we contributed to improving the accuracy of the existing pre-trained CNN-based style transfer method.

As implemented in [13], we composed the network with pre-trained model VGG-19, but reckoned in a different way for style and content feature, respectively. In the case of the content feature, semantic information is more important than the pixel unit information. Thus, we extracted the content feature C^l in the deepest layer of CNN, which retains only semantic information. For the style feature S^l, as introduced in [8], we calculated the correlation between feature map F^l by using the gram matrix (1).

$$G^l_{ij} = \sum_k F^l_{ik} F^l_{jk} \tag{1}$$

Gram matrix G^l_{ij} is defined as the inner product of feature maps i, j in layer l. The content and style feature that we obtained is utilized for loss calculation.

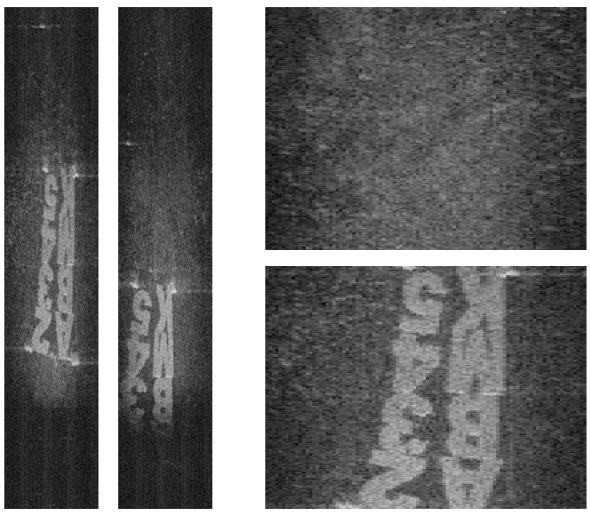


Fig. 3. The left two images are raw sonar images captured with DIDSON. The object is located in the middle of the image, and noise exists at the top and bottom only. The right two images are zoomed views of noise and object, respectively.

C. Image Generation with Loss Minimization

With the style and content features introduced above, we defined two different losses $L_{content}$ and L_{style} in accord with the concept described in [13]. The first is the style loss L_{style}; it compares the style feature between the input style and the generated image. When the style loss is minimized, the randomly generated image describes the similar style with the given style. (2) and (3) explain the style loss fuctions. Gram matrix of input style image S^l and generated image G^l are the key variables in style loss. N_l is the number of feature maps, and M_l is a size of feature map in layer l. We weighted with w_l for E_l, to control the leverage of each layer.

$$E_l = \frac{1}{N_l^2 M_l^2} \sum_{i,j} (G^l_{ij} - S^l_{ij})^2 \tag{2}$$

$$L_{style} = \sum_l w_l E_l \tag{3}$$

The second is a content loss $L_{content}$ that compares the content feature between the input content and generated image. We are able to obtain the input content in the generated image if the content loss is minimized. C^l is a feature map of the input content image from the layer l of VGG-19, F^l is a feature map of the generated image. We configured the layer value l as five in our experiment that is the deepest layer of VGG-19.

$$L_{content} = \sum_{i,j} (F^l_{ij} - C^l_{ij})^2 \tag{4}$$

The object of the entire network is to minimize both the style and content loss. However, the ratio between the content and style loss must be controlled as it could result in an overfitting problem. If one of the losses is dominant, the other loss will be neglected.

$$L_{total} = \alpha L_{style} + \beta L_{content} \tag{5}$$

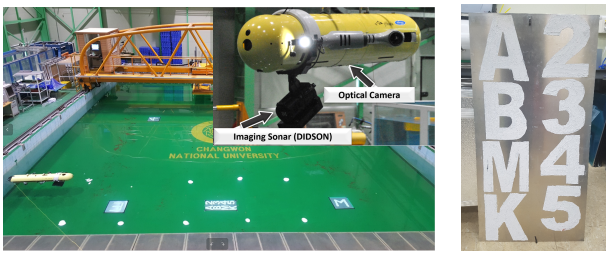
Thus, we defined total loss (5) to control the balance between style and content loss.

IV. Result

To evaluate the proposed method, an opti-acoustic stereo imaging vehicle is utilized. Using the obtained image, we infer image registration performance with cosine similarity and feature matching. The style of the generated image is similar to the optic camera image while containing the object captured with sonar.

A. Experimental Setup

Our experiment was conducted in a water tank with an AUV illustrated in Fig. 4(a). Dual frequency IDentification SONar (DIDSON) is attached to the front of the vehicle, and the optic camera is installed on the bottom of the vehicle facing down. DIDSON generates a 96×512-sized gray-scale sonar image. Simultaneously, the underwater optic camera captures a 1024×1024-sized RGB image. The target object of our experiment is Fig. 4(b), composed of numbers and

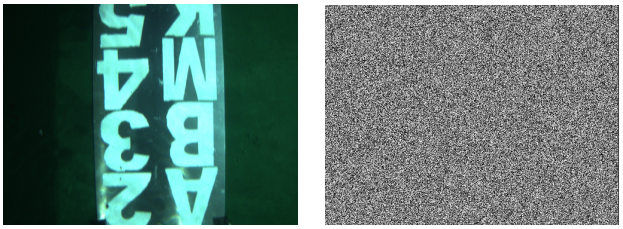


(a) Water tank and AUV (b) Target

Fig. 4. Experiment environment and target for opti-acoustic sensors. For AUV, DIDSON is linked for the front view, and the optical camera takes the bottom view.

letters. AUV pass through the water surface, and capture the target with both DIDSON and optical camera.

B. Neural Style Transfer



(a) Optic image (b) Noise image

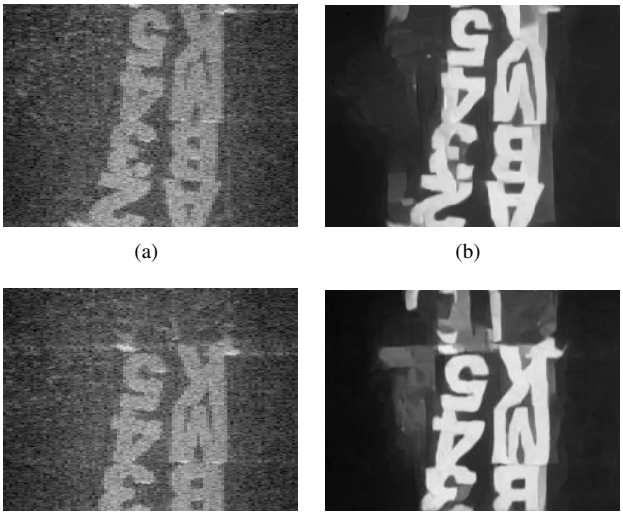
Fig. 5. Input images for network, style reference and randomly generated noise.

Our style reference is an optic image as shown in Fig. 5(a), acquired with AUV. For image generation, initial image value is produced as Fig. 5(b). Through the network, we could generate the result image depicted in Fig. 6(b) and Fig. 6(d). Since the noise of the sonar image is eliminated, the visibility of the target is increased.

C. Feature Matching Evaluation

We evaluated our style-transferred image both quantitatively and qualitatively. To validate whether the conversion was successful, cosine similarity calculation was utilized for a quantitative baseline. Specifically, extracted vectors are from the layers of VGG-19, and similarity is calculated using all pixels in the image for global comparison. Fig. 7 is the comparison result. The cosine similarity ranges from 0 to 1; the experimental group's value close to 1 signifies similarity to the control group.

Our feature matching result is shown in Fig. 8. We utilized SIFT as a feature detector, and feature matching is conducted according to Fast Library for Approximate Nearest Neighbors (FLANN) method. Fig. 8(a) and Fig. 8(c) present the entire matches with the original image. Only five feature points are matched between the optic and sonar images, yet, 25 matches occurred in the generated image. After the feature matching, we eliminated the matching error with random sample consensus (RANSAC). As we can observe in Fig. 8(b) and Fig. 8(d), stable feature matching is achieved



(a) (b)

(c) (d)

Fig. 6. (a), (c) are the original sonar images, and (b), (d) are the result images from the style conversion network. Noise is reduced, and the target became more vivid.

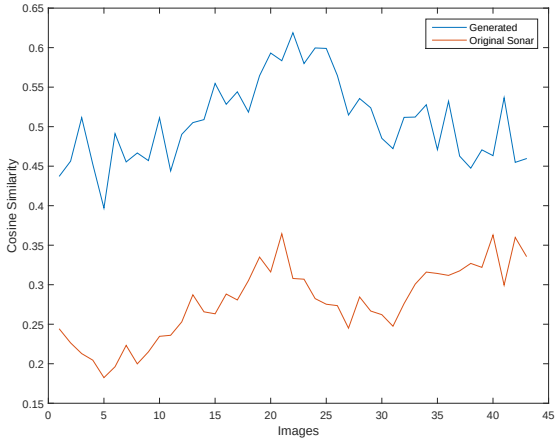


Fig. 7. Cosine similarity with optic image. Generated images provide enough higher value than original sonar images.

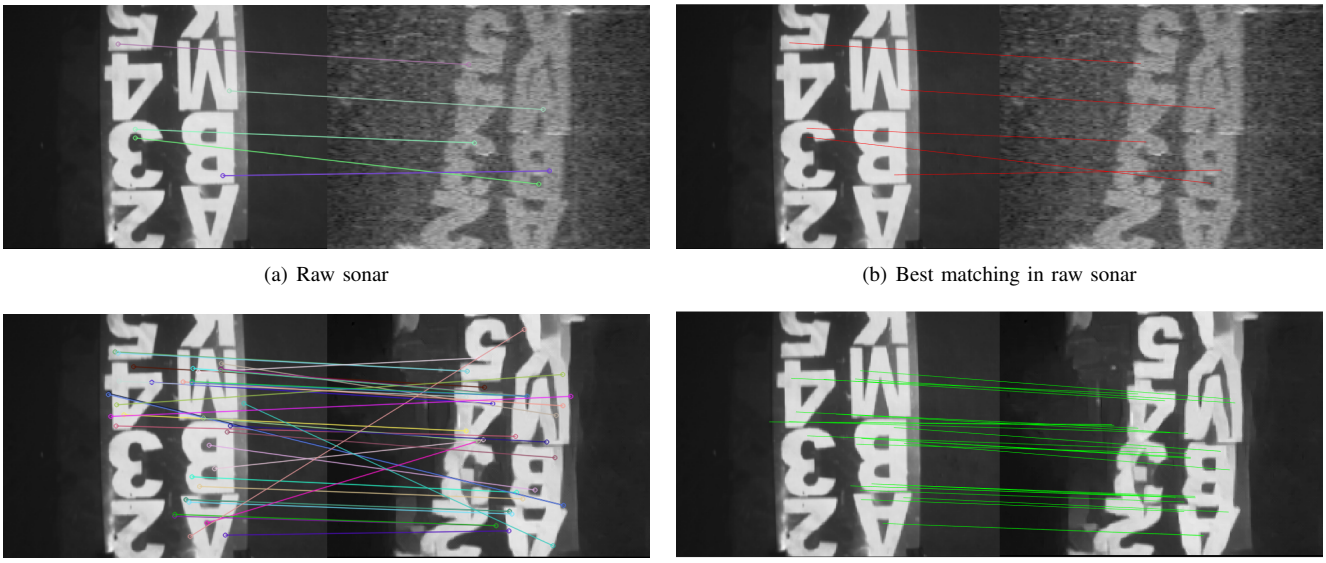
after the style transferring. Summary of the data comparison is interpreted in Table I

TABLE I
COSINE SIMILARITY WITH ORIGINAL OPTIC IMAGE

Image	Raw Sonar	Styled Sonar
Cosine Similarity	0.277	0.517
Matched feature	6	35
Precise matching	2	25

V. Discussion & Conclusion

In this paper, we presented the method to apply neural style transfer to opti-acoustic image conversion. As expected, sonar image conversion into an optic camera style was successful according to both aesthetic and analytic standards. Although the original sonar image has featureless characteristics, it is evident that the generated image depicts robust



(a) Raw sonar

(b) Best matching in raw sonar

(c) Styled sonar

(d) Best matching in styled sonar

Fig. 8. Feature matching results with FLANN. Converted image contains more feature points than raw sonar image.

feature extraction. However, two problems were detected in the style transfer method. As depicted in Fig. 6(d), the upper part of the image is filled with an acoustic object, although space is empty in the original image. This complication occurred due to the noise-to-object transference. We plan to enhance our method to prevent unreasonable transformations. Secondly, we need to re-train the network with every new input optical image. Training does not take a long time as the style transfer network is not complicated; however, it has to be improved with other methods.

For future work, we will obtain more opti-acoustic datasets to generalize our method. With a complex underwater situation, we will acquire an image that various objects are gathered in one scene. Neural network based object detection will be added in the image pre-processing step to eliminate noise part and crop the object as box form.

ACKNOWLEDGMENT

This research was supported by a grand from Endowment Project of "Development of Basic Technologies of 3D Object Reconstruction and Robot-Manipulator Motion Compensation Control for Autonomous Underwater Working" funded by Korea Research Institute of Ships and Ocean Engineering (PES3170)

REFERENCES

[1] Y. Cho, Y.-S. Shin, and A. Kim, "Online depth estimation and application to underwater image dehazing," in *OCEANS 2016 MTS/IEEE Monterey*. IEEE, 2016, pp. 1–7.

[2] Y. Cho and A. Kim, "Visibility enhancement for underwater visual slam based on underwater light scattering model," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2017, pp. 710–717.

[3] J. Li, K. A. Skinner, R. M. Eustice, and M. Johnson-Roberson, "Watergan: unsupervised generative network to enable real-time color correction of monocular underwater images," *IEEE Robotics and Automation Letters*, vol. 3, no. 1, pp. 387–394, 2018.

[4] N. Hurtós, Y. Petillot, J. Salvi *et al.*, "Fourier-based registrations for two-dimensional forward-looking sonar image mosaicing," in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*. Ieee, 2012, pp. 5298–5305.

[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] S. Negahdaripour, H. Sekkati, and H. Pirsiavash, "Opti-acoustic stereo imaging: On system calibration and 3-d target reconstruction," *IEEE Transactions on image processing*, vol. 18, no. 6, pp. 1203–1214, 2009.

[7] D.-H. Gwon, Y.-S. Shin, Y. Kim, A. Kim, Y. Lee, and H.-T. Choi, "Nontemporal relative pose estimation for opti-acoustic bundle adjustment," in *OCEANS 2016 MTS/IEEE Monterey*. IEEE, 2016, pp. 1–5.

[8] L. Gatys, A. S. Ecker, and M. Bethge, "Texture synthesis using convolutional neural networks," in *Advances in neural information processing systems*, 2015, pp. 262–270.

[9] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[10] L. A. Gatys, A. S. Ecker, and M. Bethge, "A neural algorithm of artistic style," *arXiv preprint arXiv:1508.06576*, 2015.

[11] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.

[12] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2223–2232.

[13] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2414–2423.