IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received October 6, 2020, accepted October 12, 2020, date of publication October 15, 2020, date of current version October 28, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3031297

Bat-G2 Net: Bat-Inspired Graphical Visualization Network Guided by Radiated Ultrasonic Call

SEOHYEON KIM, (Member, IEEE), GUNPIL HWANG, AND HYEON-MIN BAE, (Member, IEEE)
Korea Advanced Institute of Science and Technology, Daejeon 34141, South Korea
Corresponding author: Hyeon-Min Bae (hmbae@kaist.ac.kr)

ABSTRACT In this paper, a noise-immune Bat-inspired Graphical visualization network Guided by the radiated ultrasonic call (Bat-G2 net) that can reconstruct 3D shapes of a target from ultrasonic echoes is presented. The Bat-G2 net achieves noise-resiliency by emulating bat's auditory system that processes echoes along with the highly correlated radiated ultrasonic call (RUC). In order to extract the information contained in the echoes robustly and effectively, two implementation ideas have been applied to the Bat-G2 net: (1) RUC-guided attention, and (2) non-local attention. The Bat-G2 net is trained with ECHO-4CH dataset acquired by a custom-made Bat-I sensor. Noise-resistant property of the Bat-G2 net is demonstrated by comparing the reconstructed images with those from current state-of-the-art ultrasonic image reconstruction network under low SNR conditions. This study clearly demonstrates the implementation feasibility of the new modality of 'seeing by hearing' in practical environments.

INDEX TERMS 3D reconstruction, biologically inspired vision, deep learning: applications, methodology, and theory, graphics, vision applications and systems, vision for robotics, visual reasoning.

I. INTRODUCTION

Recently, sensors and information processing technologies have led to the growth of unmanned systems (UMS) such as drones, autonomous vehicles, and robots. For UMS to reach a fully autonomous level that does not require human intervention, the sensors employed in UMS must be able to detect surroundings irrespective of dynamic environmental obstacles such as weather, temperature, and noise. To reach this goal, UMS is typically designed to use a combination of sensors (RGB-D cameras, RADARs, LIDARs, and ultrasonic sensors), which are complementary to each other. RGB-D cameras and LIDARs provide sophisticated visual information. However, their accuracy and visibility are susceptible to deterioration depending on the environment and weather conditions. While, RADARs and conventional ultrasonic sensors, which measure time-of-flight from a target object, are robust to the environmental interferences, they provide only low-resolution ranging information [1], [2]. In conclusion, there is a growing consensus on the need for 3D imaging sensors that are virtually unaffected by environmental conditions.

The associate editor coordinating the review of this manuscript and approving it for publication was Peter Peer.

Lately, there has been a promising attempt to meet this demand. A high-resolution ultrasound 3D imaging system referred to as a Bat-G network has been proposed, which employs a feed-forward neural network emulating the echolocation mechanism of a live bat [3]. The Bat-G net is designed to reconstruct the 3D representation of a target object from hyperbolic frequency-modulated (HFM) echoes reflected from the target object. In case the object has the small sonar cross section (SCS) or has been measured in a noisy environment, the signal-to-noise ratio (SNR) of the received echo drops below 0 dB. In such low SNR condition, the imaging accuracy can be compromised significantly since the Bat-G net cannot distinguish between the received echo and noise. Since it is a challenge to ensure high SNR sensor data in practical unmanned system environments, it is essential to classify the sensory input data into echoes that plays a crucial role in 3D imaging, and unnecessary signals, for robust 3D imaging under low SNR conditions.

In this paper, we propose a robust noise-immune ultrasound 3D imaging network referred to as a Bat-inspired Graphical visualization network Guided by the radiated ultrasonic call (Bat-G2 net) that visualizes 3D spaces under low SNR conditions like live bats. Such immunity is achieved by actively utilizing a radiated ultrasonic call (RUC) during the

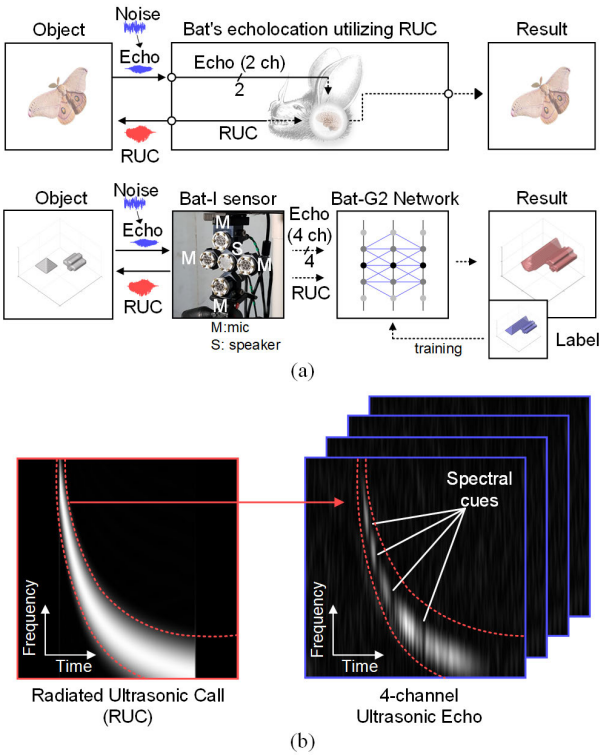

FIGURE 1. (a) Bat's echolocation utilizing RUC and 3D ultrasound imaging system utilizing RUC. (b) Radiated Ultrasonic Call (RUC) and ultrasonic echo (4 channels).

echo processing as a prior knowledge, as shown in Fig. 1(a). Live bats utilize the ultrasound echoes reflected from target objects to localize and discriminate objects, which is called echolocation [4]. Bats can discriminate 10-15 ns fine delay even in an acoustically harsh environment where dozens of bats are simultaneously emitting ultrasonic calls [4]. Such exceptional sensing capability is attributed to the bat's auditory system that utilizes the fact that informative received echo is highly correlated with the RUC, as shown in Fig. 1(b) [5]. From this, we can deduce that the proper use of RUC is the key to building a noise-resilient ultrasonic 3D imaging system, as shown in Fig. 1(a). In order to utilize the information contained in RUC effectively, two implementation ideas have been applied to the Bat-G2 net: *(1) RUC-guided attention*, and *(2) Non-local attention*. *(1) RUC-guided attention* - Conventional self-attention creates attention maps from the extraction of its features. In case the SNR of the received signal is low, the generated self-attention map from the received ultrasound signal is inadequate to distinguish the echoes from target objects and the pure noise. To overcome this problem, the proposed network employs a RUC-guided attention method, generating learnable attention kernels by employing both reflected echoes and high SNR RUC, as shown in Fig. 1(a). *(2) Non-local attention* - Because the shape and ranging information of target objects are encoded in the spectral pattern over a wide frequency range in the received ultrasonic signal, as shown in Fig. 1(b), a non-local attention module is adopted in order to decode such a pattern existing at non-local spatial locations in the sensory input image.

II. RELATED WORK

Airborne ultrasonic sensors have been one of the leading range detection sensors for decades due to their simplicity. These sensors calculate the distance to an object by emitting a single frequency ultrasonic signal and measuring the time-of-flight (TOF) of the echo reflected from the object. There have been approaches to localize/classify a target object or reconstruct the shape of an object, as shown in Table 1. Representative 3D localization strategies include calculating the TOF difference between two pairs of microphones [6] and concentrating signals to a designated direction using beamforming (BF) techniques [7]. In another line of research, classification of a target object has been achieved by employing the angle/distance between the 3D sensor array and an object as the classification parameters [8] or by applying the principal component analysis (PCA) method to 16 TOF vectors (4 TXs and 4 RXs) [9]. However, because such techniques relies on a lookup table, only simple objects such as planes, corners, and edges are discriminated. On the other hand, [10] makes an attempt to classify the cubes and tetrahedrons by analyzing the spectrum of echoes employing neural networks (NN). Nevertheless, such an effort did not exploit the full potential of the NN approach due to limited datasets and primitive NN structure. In addition to 3D localization and classification of target objects, there have been other attempts to reconstruct the 3D shape of objects from the received echoes by solving ill-posed inverse problems. Such attempts include BF [11] and holography [12] methods using a large number of TRX arrays. An approach has been made to reconstruct cuboids in sparse scenes using Compressive Sensing (CS), a subset of the inverse problem approach, in the simulation domain [13]. However, these inverse problem approaches demand rigorous calibration and heavy computational power and time to process the incoming data from a large number of arrays. In [3], a feed-forward NN emulating the auditory neural network of bats reconstructs the 3D representation of various objects. However, such NN is not readily applicable to UMS because the reconstruction performance is susceptible to significant deterioration under low-SNR environmental conditions. In this paper, a robust noise-resilient Bat-G2 net visualizing 3D spaces from the ultrasonic echoes is proposed. The proposed network has been designed with the inspiration from remarkable RUC-involved imaging capability of live bats in acoustically challenging environments.

III. PRELIMINARIES

In order to understand the 3D spatial perception mechanism of live bats that represent 3D space from an ultrasonic echo, it is necessary to understand the bat's auditory system, which consists of three main elements: *(1) cochlear* and *(2) temporal cue analysis* and *(3) spectral cue analysis* block. *(1) Cochlear* - The basilar membrane in the cochlear of a bat with frequency selectivity according to its location can be

TABLE 1. Summary of Related Works.

	[6]	[7]	[8]	[9]	[10]	[11]	[12]	[13]	[3]	**This Work**
3D localization	✓	✓	✓	✓	×	✓	✓	✓	✓	✓
Classification	×	×	✓	✓	✓	✓	✓	✓	✓	✓
Reconstruction	×	×	×	×	×	✓	✓	✓	✓	✓
TX / RX	1/4	1/32	3/3	4/4	1/1	1/400	1/64	5/3	1/4	**1/4**
Measurement	✓	✓	✓	✓	✓	✓	✓	×	✓	✓
Noise Consideration	×	×	×	×	×	×	×	×	×	✓
Method[a]	T	BF	T, AC	T, PCA	NN, BM	SA, BF	NN, HG	T, CS	T, NN, BM	**T, NN, BM**

NOTE: [a] T: Time difference of arrival, BF: Beamforming, SC: Spectral Cues, BM: Biomimetics, AC: Angle Change, PCA: Principal-Component-Analysis, NN: Neural Network, SA: Synthetic Aperture, HG: Holography, CS: Compressive Sensing

represented by an array of band-pass filters (BPFs) followed by a half-wave rectifier and a low-pass filter (LPF) at the output of each BPF [14], [15]. As a result, the filter banks decompose the emitted/received sound signal into band-pass filtered signals according to the frequencies. Then, the signal intensity (or power) of each frequency channel is extracted by a subsequent rectifier and an LPF. Such processes convert the acoustic time-domain signal into a time-frequency representation that is similar to the spectrogram. *(2) Temporal cue analysis (TCA)* - Elapsed time between the emitted sound signal and its echoes are measured by the TCA block. The elapsed time is calculated by delay-tuned neurons and coincidence detection neurons carrying out cross correlation function of the emitted and the received sound signal. *(3) Spectral cue analysis (SCA)* - The TCA block mechanism cannot discriminate the fine delay produced by overlapping echoes reflected from two nearby glints. These fine delays are deciphered by the SCA block analyzing spectral cues such as notches and nulls [14], [16], [17], [19], [21]. Since a target object is acoustically composed of several glints and reflective surfaces, the spectral cues of a target object consist of the summation of echoes reflected from several glints and surfaces [22]–[26]. In other words, the shape of an object is represented by its spectral fingerprint [15], [27], [28]. Consequently, sophisticated spectral pattern recognition of the received ultrasonic echoes is the key to 'Seeing by hearing'.

IV. APPROACH

This section describes two key design schemes resulting in noise-resilient NN-based echolocation system: a radiated ultrasonic call (RUC) guided attention method and a non-local operation.

A. RUC GUIDED ATTENTION METHOD

Bats live as groups in habitats with significant reverberation, which exposes them to jamming effect by voices of other bats. Such a severe circumstance makes it quite challenging for bats to perceive their surroundings using ultrasound. In order to survive in this acoustically harsh environment, bats establish an auditory system that can detect preys in the order of centimeters by actively utilizing the fact that the received echoes are especially correlated with the RUC. Therefore, the mechanism effectively processing

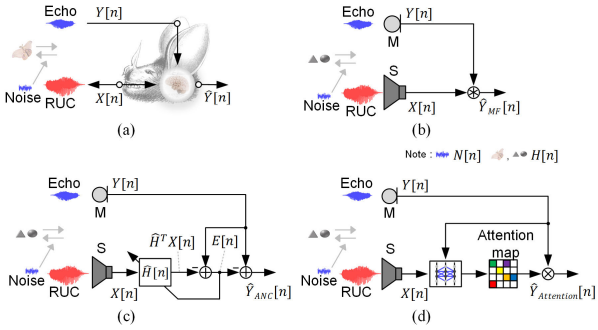

FIGURE 2. (a) Bat's auditory system utilizing RUC. (b) Matched filter utilizing RUC. (c) Adaptive noise cancellation using RUC. (d) RUC guided attention method.

echoes along with correlated RUC is critical for the bat's extraordinary performance in extremely noisy environments, as shown in Fig. 2(a). In order to discriminate the echo from noisy received signals, a **3) RUC GUIDED ATTENTION METHOD** is proposed and is compared with two conventional noise canceling methods: **1) MATCHED FILTER** and **2) ADAPTIVE NOISE CANCELLATION**.

1) MATCHED FILTER

A technique applied to human-made systems such as RADAR or SONAR detects the echo signal in the noisy received signal by calculating the correlation of the received signal with RUC, as described in Eq.(1)-Eq.(3) (see Fig. 2(b)).

$$Y[n] = H^T X[n] + N[n], \tag{1}$$

$$\hat{Y}_{MF}[n] = (H^T X[n] + N[n]) * X[n], \tag{2}$$

$$\hat{Y}_{MF}[n] = H^T R_{XX}[n] + R_{XN}[n], \tag{3}$$

where $X[n]$ is RUC, H^T is the impulse response of target object, $N[n]$ is the additive stochastic noise, $Y[n]$ is the noisy received signal, $\hat{Y}_{MF}[n]$ is the output of the matched filter guided by correlating with RUC $X[n]$, as described in Eq.(2), $R_{XX}[n]$ is the auto-correlation function of the signal and $R_{XN}[n]$ is the cross-correlation function of the noise and signal. Matched filter technique on the basis of direct correlation with RUC achieves outstanding detection of the echo signal in an environment with white noise ($\because R_{XX}[n] \gg R_{XN}[n]$). However, in case the environmental noise is correlated with the RUC ($\because R_{XX}[n] \not\gg R_{XN}[n]$) such as reverberation (see Eq.(3)), the matched filter method fails to detect echoes effectively.

2) ADAPTIVE NOISE CANCELLATION (ANC)

An adaptive filtering technique supresses various interferences as well as the noise effectively by subtracting an adaptively approximated noise using an optimization algorithm (e.g. Least Mean Squares (LMS)) from the noisy received signal, as shown in Fig. 2(c). The noise reduction process of the ANC employing the RUC can be simply described as follows: Eq.(4)-Eq.(6),

$$Y[n] = H^T X[n] + N[n], \tag{4}$$

$$E[n] = Y[n] - Y'[n] = Y[n] - \hat{H}^T X[n] = N'[n], \tag{5}$$

$$\hat{Y}_{ANC}[n] = Y[n] - E[n] = H^T X[n] + N[n] - N'[n], \tag{6}$$

where $X[n]$ is RUC, H^T is the impulse response of target object, and $N[n]$ is the additive stochastic noise. The parameter of the filter $\hat{H}^T$ is adaptively adjusted to minimize the error signal $E[n]$. The approximated noise $N'[n]$ is estimated by subtracting the RUC-guided adaptive filter output $\hat{H}^T X[n]$ from the noisy received signal $Y[n]$, as shown in Eq.(5). Consequently, the noise reduction is achieved by subtracting adaptively approximated noise $N'[n]$ from noisy received signal $Y[n]$, as shown in Eq.(6). However, this ANC technique based on a linear filter does not lead to the full potential of RUC guidance since the ANC technique is unable to suppress nonlinearity or non-stationary noise/interference effectively.

3) RUC GUIDED ATTENTION METHOD

In order to compartmentalize the signal surrounded by various noise/interference, we adopted a neural network (NN) known to operate as a universal approximator [29] for the RUC guidance, as shown in Fig. 2(d). By introducing non-linearity via an activation function, NN has extraordinary capability to represent a wide variety of functions, which makes it effective in handling the nonlinearity or non-stationary noise. The structure of the proposed NN to focus on the sensory signals is inspired by the ''attention'' concept distinguishing the informative regions to be emphasized. In recent years, attention mechanisms have shown promising results in a variety of computer vision tasks such as image classification [30]–[32], object detection [33], [34], image captioning [35]–[37] and visual question answering [35], [38]. Such conventional self-attention networks are implemented by applying an adaptively trained attention map from the extraction of its own features [39]. However, applying the attention concept directly to a network handling low-SNR sensory inputs cause unsatisfactory results. Therefore, the proposed RUC-guided attention network employs a learnable kernel relying not only on sensory input but also on RUC with high SNR, as shown in Fig. 2(d).

B. NON-LOCAL OPERATION

The shape/location of an object is encoded in a spectral fingerprint (see section III.) that appears at non-local spatial locations in the echo spectrogram, as shown in Fig. 1(b). For reliable long-range contextual feature extraction from the ultrasonic echoes, we have adopted a non-local attention mechanism. The non-local attention method calculates the correlation between all pixels in input features so that a pixel in each position in the feature map is represented with all other ones. By this means, non-local operation effectively captures long-range contextual features placed even in non-local spatial locations [40]. Eventually, the guided non-local attentional module (GAM) has been introduced by combining the non-local operation with the RUC-guided attention approach.

V. ARCHITECTURE OF PROPOSED BAT-G2 NETWORK

In this section, the architecture of the proposed noise-resilient Bat-G2 net reconstructing a 3D representation of a target object from four-channel ultrasound echoes is presented. The architecture is described in two perspectives: (A) an encoder-decoder structure which emulates the central auditory pathway of bats that individually process the temporal/spectral features and (B) a bat-inspired guided non-local attention module (GAM) combining RUC-guided attention and non-local operation for noise-immune echo decryption.

A. ENCODER-DECODER

The encoder-decoder architecture literally consists of an encoder that extracts information from the received echoes and a decoder that projects the output data of the encoder in low dimensional manifold into the 2D depth image, as shown in Fig. 3. The specific structure of each part is as follows.

1) ENCODER

A Bat's central auditory system is composed of neurons that are particularly sensitive to either temporal- or spectral-domain information [3]. Such neurons form a cluster of neurons, referred to as a nucleus and each nucleus processes domain-specific cues separately depending on the characteristics of comprising neurons. Bat's extraordinary perception ability is originated from such separated processing of temporal and spectral components in the bat's central auditory system [41]. In order to achieve such performance, the proposed encoder is implemented with two separated pathways specialized for extracting temporal- or spectral-domain features from the input spectrogram. In order to feed the appropriate domain-specific input to each path, the recorded sound signals on each channel (right, left, up, and down) are transformed into two high-resolution spectrograms in frequency/time using the short-time Fourier transform (STFT) with long/short window (LW/SW). The spectrograms are fed to the GAM, which will be covered in the following subsection in detail. The output features of the GAM passes through a residual block (RB)-2 that is composed of a convolutional path (three 3×3.2D convolution layers with a 2×2 max-pooling layer) and a residual path (a 1×1.2D convolution layers with a 2×2 max-pooling layer). The RB-2 elicits the spectral/temporal cues necessary for the reconstruction of shape/position from each channel's spectrogram. In addition, the RB-2's output feature maps are concatenated and then fed to a 1×1 convolution layer with a successive RB-4 layer in order to extract the hidden clues in the time-of-flight (TOF) differences in reflected signals between the channels.

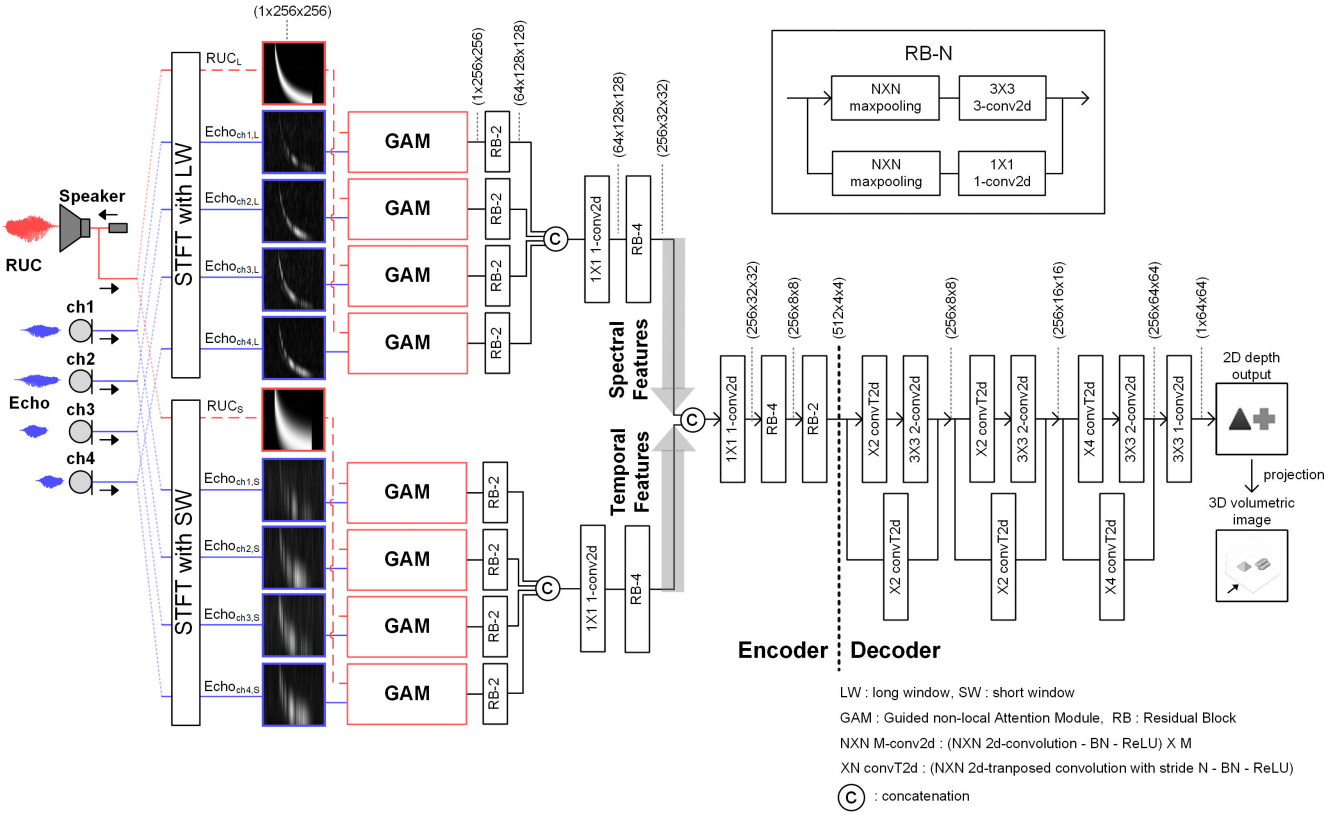

FIGURE 3. **Overall architecture of the proposed Bat-G2 net.**

Finally, the entire products of spectral- and temporal-pathways are integrated and encoded in a form suitable for the decoding process using series connected RB-4 and RB-2 modules.

2) DECODER

An inverse rendering decoder transforms the output features of the encoder in the low dimensional manifold to the 2D depth image in $\mathbb{R}^{64\times64}$ vector space (see Fig. 3). Three decoding residual blocks are applied to 4×4 pixel inputs encoded with 512 feature maps. Each residual block is composed of one convolution transpose (or deconvolution) layer (stride-2 2×2 or stride-4 4×4 kernels) and two convolution layers (3×3 kernels and same padding with batch normalization (BN) and rectified linear unit (ReLU)). To convert 256 output feature maps of the residual block into the desired representation, a 3×3 convolution layer is inserted to the final layer. Please refer to Fig. 3 for the detailed composition of each layer.

B. GUIDED NON-LOCAL ATTENTION MODULE

The spectral/temporal cues for the shape/location of a target object appears at a non-local location in the echo spectrogram (see Fig. 1(b)). In order to decode such cues effectively, an RUC-guided non-local attention module (GAM) is implemented. In a typical neural network, the spatial information contained in the feature map is flattened as the layer depth increases, while the semantic information is enhanced. Therefore, the GAM is placed at the forefront of the Bat-G2 net to extract spatial information embedding the spectral/temporal cues before the decrease of the resolution. However, large input feature size resulting from such placement requires tremendous computational resources due to high complexity $O(HW \times HW)$ of the spatial attention map, where H and W are the height and the width of the feature map, respectively. In order to reduce the computational cost, the core of the GAM is designed to be encapsulated in sub-pixel sampling layers including down pixel-shuffle block and up pixel-shuffle block, so that the core attention method operates at low feature dimensions. In addition, such sub-pixel sampling method preserves the spatial information because the entire feature map is subsampled at a specific scale ratio r, as shown in Fig. 4(b) [42]. The feature maps of the RUC (Fr) and the echo signal (Fe) are subsampled by the down pixel-shuffle block and then embedded into the low dimensional manifolds ($(f(F_r^D)$ and $g(F_e^D))$ through two 1×1 convolutions f and g, respectively. The GAM computes the attention map at a pixel as a weighted sum of two embedded features $f(F_r^D)$ and $g(F_e^D)$ at all pixels as done in non-local NN [40]. For ease of implementation using a network platform, the attention map is implemented using the Gaussian function $e^{x^T x}$ with a normalization factor $C(f(F_r^D), g(F_e^D))$,

$$Atten\_Map(F_r^D, F_e^D) = \frac{1}{C(f(F_r^D), g(F_e^D))} e^{f(F_r^D)^T g(F_e^D)}. \quad (7)$$

Assuming the normalization factor $C(f(F_r^D), g(F_e^D)) = \sum_{\forall j} e^{(f(F_r^D)_i^T, g(F_e^D)_j)}$, where i is the index of a position in the feature map and j is the index enumerating every possible

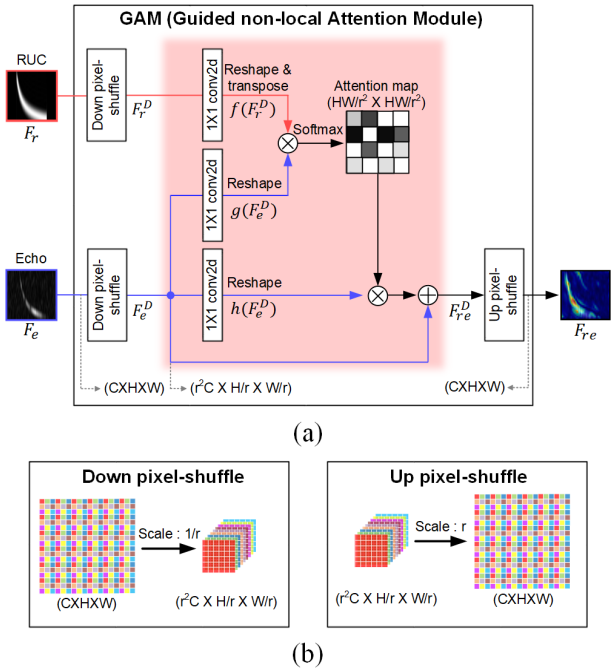

FIGURE 4. **(a) Structure of Guided non-local Attention Module (GAM). (b) Operation of sub-pixel sampling (down pixel-shuffle and up pixel-shuffle).**

positions, the attention map is equal to the softmax function. Then, the feature map of the RUC-guided echo in the subsampled dimension, F_{re}^D is

$$\begin{aligned} F_{re}^D &= Atten\_Map(F_r^D, F_e^D)h(F_e^D) + F_e^D \\ &= softmax(f(F_r^D)^T g(F_e^D))h(F_e^D) + F_e^D, \end{aligned} \tag{8}$$

where $h(F_e^D)$ is the embedded feature of the echo signal applying the attention. For training efficiency in practice, the subsampled feature of an echo F_e^D is added as identity mapping [43]. Finally, F_{re}^D is restored to the original feature dimension by using the up pixel-shuffle layer and the RUC-guided echo F_{re} can be written by

$$\begin{aligned} F_{re} &= (F_{re}^D)^U \\ &= (softmax(f(F_r^D)^T g(F_e^D))h(F_e^D) + F_e^D)^U, \end{aligned} \tag{9}$$

where $(F_{re}^D)^U$ is the up pixel-shuffle operation of the feature F_{re}^D.

VI. EXPERIMENTS

A. DATASET AND TRAINING

Performance of the proposed Bat-G2 net is evaluated using the ECHO-4CH dataset [3]. ECHO-4CH consists of the measurements of 16.2k geometric object configurations using custom-made Bat-inspired imaging (Bat-I) sensor (see Fig. 1). The Bat-I sensor emits a hyperbolic frequency-modulated (FM) chirp in the frequency range of 20-120kHz with the duration of 6ms and records echoes reflected from target objects. Finally, the measured signal is converted into two high-resolution spectrograms in frequency/time using the short-time Fourier transform (STFT) with long/short hamming window ($133\mu s/33\mu s$ window size with $90\mu s/22\mu s$ overlap). Therefore, we have built a large ECHO-4CH dataset (49k data for training and 2.6k data for evaluation). Each data is composed of eight spectrograms (256^2 grayscale image) and one 2D ground-truth label (64^2 pixels) which is projected from 3D model (detailed in [3]).

The Bat-G2 Net is trained by employing a supervised learning algorithm. The network is continuously fed with 49k training data randomly selected from the ECHO-4CH dataset with the batch size of 16. Adam optimizer with a learning rate of 0.001 has been adopted for better convergence. In the following experiments, we trained a network iteratively for 200 epochs (613k steps) on a GTX 1080 Ti GPU and a Threadripper 1900X CPU using a Mean-Square Error (MSE) as a loss function. The total time for learning is 85 hours causing from taking 0.5 seconds/step.

B. EXPERIMENTAL RESULTS

We first conduct a qualitative analysis of the reconstruction results of the proposed Bat-G2 network and the efficacy of the GAM. Then, reconstruction performance is quantitatively evaluated based on the peak signal-to-noise ratio (PSNR) and the structural similarity (SSIM) index. In such qualitative and quantitative assessment, the current state-of-the-art ultrasonic image reconstruction neural network, Bat-G net [3], is employed as the baseline. The Bat-G2 net is evaluated with 2.6k test data from the ECHO-4CH dataset.

1) RECONSTRUCTION

We conducted three qualitative assessments to verify the performance of the Bat-G2 net: *(1) 3D reconstruction with object shape dependency*, *(2) 3D reconstruction sensitivity with respect to SNR*, and *(3) 3D reconstruction under diverse interferences*. For visual comparison, 2D output image of the Bat-G2 net is converted into volumetric 3D image using the inverse projection of 2D label image. *(1) 3D reconstruction with object shape dependency* - When a radiated ultrasonic chirp is reflected from convex surfaces of target objects, the 3D representation of the measured objects is correctly reconstructed, as shown in Fig. 5(a)-(b). Since the shape of an object is reconstructed using the reflected echoes, the visualization of an object with small or slanted reflective surface is difficult due to lack of information, as shown in Fig. 5(c)-(d). *(2) 3D reconstruction sensitivity with respect to SNR* - SNR of sensory input data varies depending on the reflected signal power and the noise level. The signal power is mainly determined by the sonar cross section (SCS) of an object and the distance between the sensor and the object, and the amount of noise is determined by the level of ambient and the electrical noise. In order to investigate the noise-immunity of the proposed Bat-G2 net, the reconstructed 3D representations were compared under various SNR conditions, as shown in Fig. 6. In case the SNR is high (see Fig. 6(a)), both baseline and the Bat-G2 net demonstrate high quality 3D image reconstruction of the measured objects. As the SNR decreases, the baseline network generates distorted 3D image

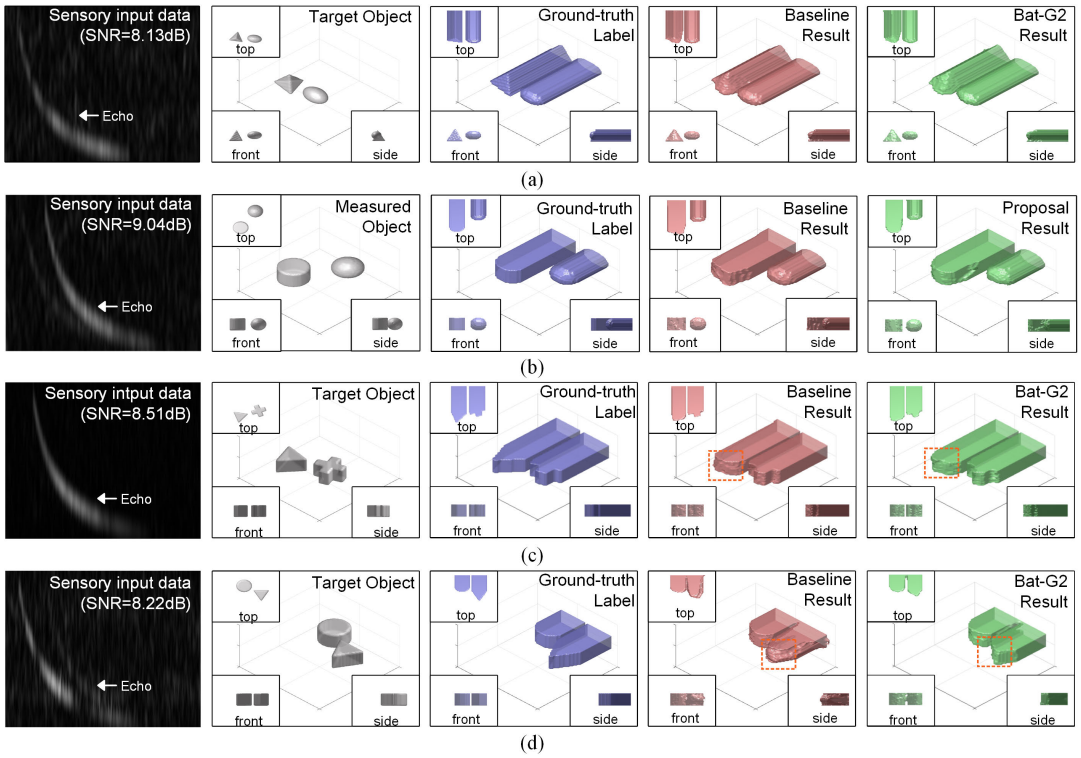

FIGURE 5. 3D reconstruction results with object shape dependency.

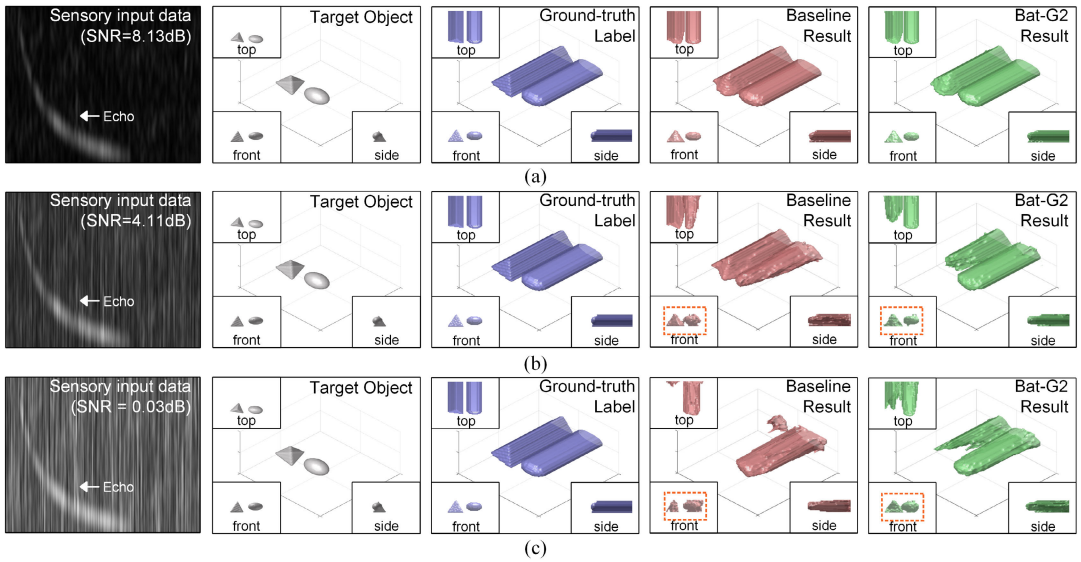

FIGURE 6. 3D reconstruction results to sensitivity with respect to SNR ((a) SNR=8.13dB, (b) SNR=4.11dB, and (c) SNR=0.03dB).

and eventually fails to reconstruct edges clearly as shown in Fig. 6(b)-(c). In contrast, the proposed Bat-G2 network stably retrieves the 3D representation of the target object under severe SNR conditions thanks to the RUC-guidance. *(3) 3D reconstruction under diverse interferences* – In an environment where multiple unmanned systems (UMS) exist, the broadcasted chirp signals from neighboring UMS become interferences. In order to verify the visualization capability of the Bat-G2 net under such environment, diverse interfering chirp signals with modified slope, duration (frequency range 20kHz-138kHz, duration 5.1ms) are co-generated together with the main chirp signal having the same magnitude [44], [45]. The 2D correlation coefficient between the interfering chirp and the main chirp has a high correlation of 0.5. In the absence of the interferences, both baseline and Bat-G2 net achieve high quality 3D reconstruction performance, as shown in Fig. 7(a). The baseline shows less clearer edge retrieval performance as compared to the Bat-G2 net when the

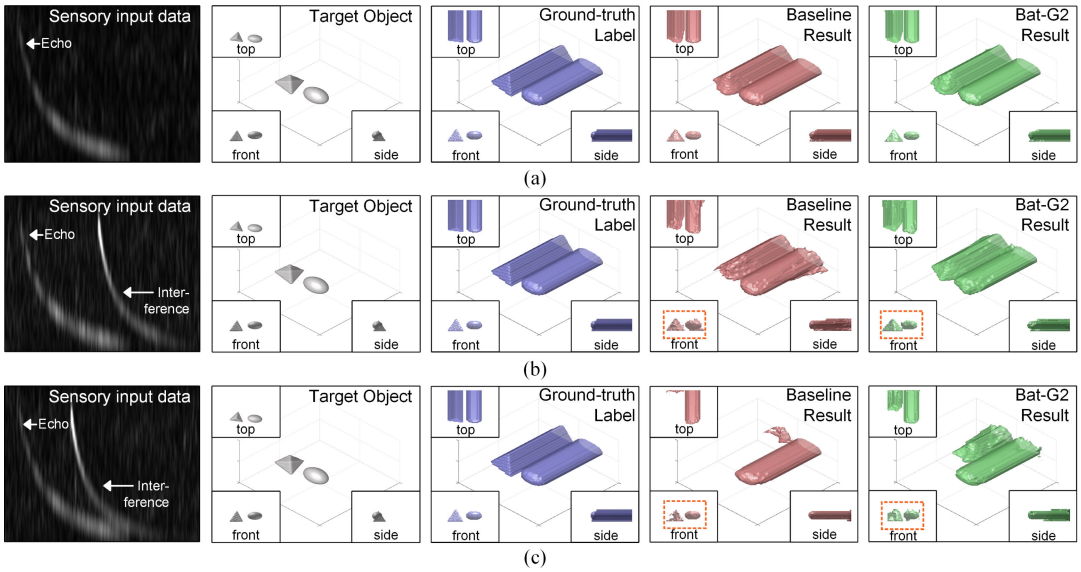

FIGURE 7. 3D reconstruction results under diverse interferences ((a) No interference, (b) Interference source is located at a distance of 2.5m in radius when the reflected echo arrives at the sensor, and (c) Interference source is located at a distance of 2.0m in radius when the reflected echo arrives at the sensor.

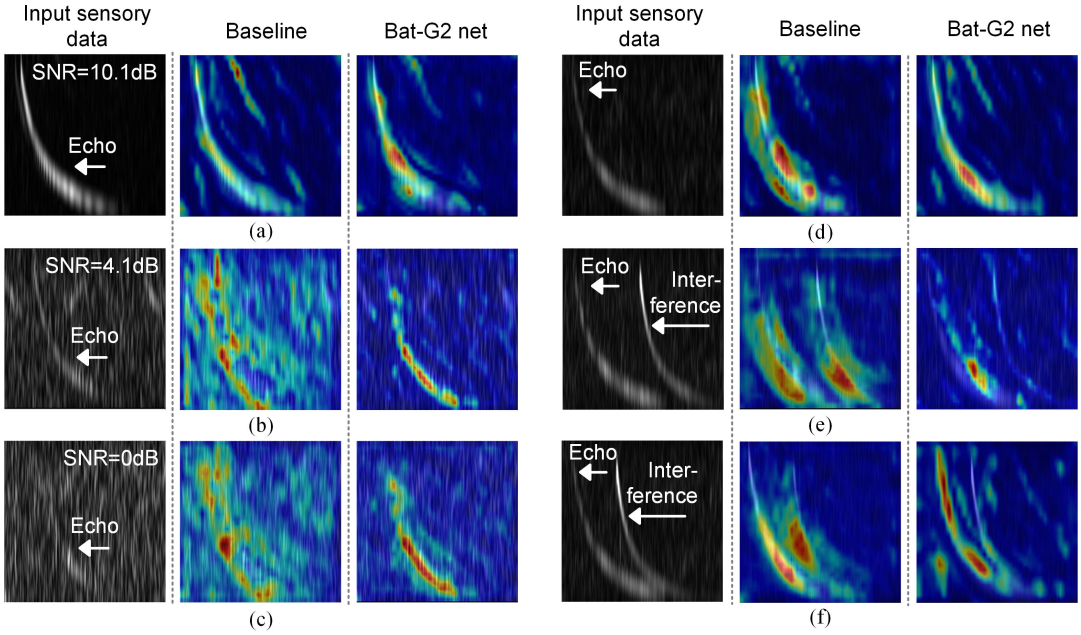

FIGURE 8. Comparison of Grad-CAM visualization results between the baseline (Bat-G net) and the proposed method (Bat-G2 net) under various SNRs when (a) SNR=10.1dB, (b) SNR=4.1dB, (c) SNR=0dB and under diverse interferences (d)-(f).

interference starts to overlap with the main echoes, as shown in Fig. 7(b). As the overlay between the interference and the main echo increases, the baseline fails to reconstruct slanted reflective surface while the Bat-G2 net maintains reasonable reconstruct performance, as shown in Fig. 7(c). These results clearly demonstrate the interference immunity of the RUC-guided approach.

2) GRAD-CAM VISUALIZATION

Grad-CAM is recently proposed as a visualization method that uses gradients to locate activated spatial regions in convolution layers [46]. These spatial regions activated in a convolution layer indicate the areas that a network considers important. This visualization allows the network designers to analyze and verify whether the networks are utilizing the features properly. *(1) Grad-CAM under various SNR* - For an input image with high SNR shown in Fig. 8(a), both the baseline and the Bat-G2 net completely focus on the hyperbolic frequency-modulated (HFM) chirp. However, as the SNR decreases, the baseline cannot distinguish the HFM chirp from noise, as such it concentrates on noise as well as HFM chirp. In contrast, the Grad-CAM on the Bat-G2 net demonstrates that it strives to focus mainly on the spectral fingerprint of the HFM chirp in the echo

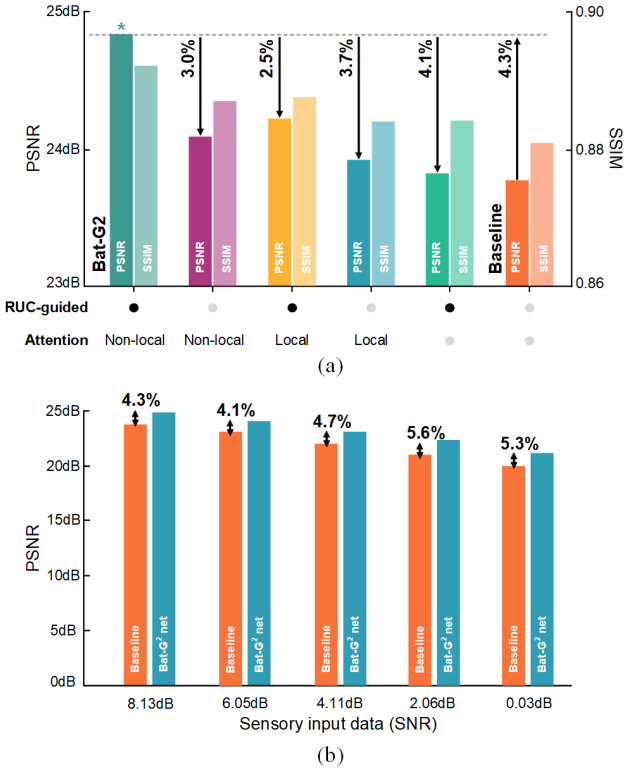

FIGURE 9. **(a) PSNR and SSIM results of the case with the RUC-guided and the case without the RUC-guided employing non-local attention, local attention, or no attention. (b) PSNR results of the baseline and the proposed Bat-G2 net under diverse SNRs.**

spectrogram, as shown in Fig. 8(b)-(c). *(2) Grad-CAM under various interferences* – In the absence of interference, both the baseline and the Bat-G2 net completely focus on the hyperbolic frequency-modulated (HFM) chirp, as shown in Fig. 8(d). As interference approaches the echo signal, the baseline focuses on both the interference and the main echoes while Bat-G2 net focuses mainly on the main echoes, as shown in Fig. 8(e)-(f). In conclusion, Grad-CAM visualization justifies the robustness of the Bat-G2 net under harsh environments.

3) QUANTITATIVE ANALYSIS

We quantitatively assessed the imaging performance of the Bat-G2 net using PSNR for the fidelity evaluation of the image and SSIM index used for the measurement of the similarity between two images (in this case, a ground-truth label and a represented image). The Bat-G2 net (RUC-guided, Non-local attention) achieved 24.84dB PSNR and 0.892 SSIM as shown in Fig. 9(a), which are 4.3% and 1.2% increase against the baseline, respectively. The parameters of the network with and without the GAM module are 21,477,889 and 21,328,001, respectively. The overhead of the GAM module is only 0.7%. When the proposed Bat-G2 network is compared with parameter-increased baseline network that has identical number of parameters to the Bat-G2 net, the Bat-G2 net shows 4.1% increase in PSNR. In addition, ablation studies are conducted to verify the efficacy of the RUC-guided method and the non-local attention approach. The contribution of the former is first assessed. By removing the RUC guide path, the reconstruction performance of the Bat-G2 net deteriorates (3.0%, 1.2%, and 0.2% drop in PSNR and 0.6%, 0.4%, and 0.4% drop in SSIM when the non-local attention, local attention, and no attention is used, respectively). This suggests that the RUC-guided approach has made a significant contribution to the extraction of meaningful features from the ultrasonic chirp. Secondly, we compared the performance of the non-local attention method with that of the local attention (implemented with a convolutional block attention module (CBAM) [46]) and the case without attention. In the cases of applying RUC-guided method, the reconstruction performance of the Bat-G2 net degrades (2.5%, 4.1% drop in PSNR and 0.5%, 0.9% drop in SSIM under local attention and no attention, respectively). Such decline demonstrates the effectiveness of the non-local attention mechanism. In addition, we validate the noise-resilient property of the Bat-G2 net as compared to the baseline under various SNR conditions, as shown in Fig. 9(b). As the SNR decreases, PSNR of both the baseline and the Bat-G2 net decreases steadily, but the Bat-G2 net outperforms the baseline. It clearly indicates the noise resilience of the RUC-guide scheme incorporated in the Bat-G2 net.

VII. CONCLUSION

In this paper, a noise-immune Bat-inspired Graphical visualization network Guided by the radiated ultrasonic call (Bat-G2 net) that can reconstruct 3D shapes of a target from ultrasonic echoes is presented. In order to decode the information contained in the radiated ultrasonic call (RUC) robustly and effectively, two implementation ideas have been applied to the Bat-G2 net: *(1) RUC-guided attention* – Attention module generating learnable attention kernels by utilizing both sensory input and high SNR RUC, and *(2) non-local attention* – A method wisely capturing a non-local spectral fingerprint of the shape/location of an object. Through a range of experiments and assessments, we have shown promising results that the Bat-G2 net maintains outstanding imaging capabilities in noisy situations. The grad-CAM visualization justifies the stable 3D reconstruction capability of the proposed network under low SNR and demonstrates the robustness to interference. Quantitative analysis based on PSNR and SSIM shows that RUC-guided approach and non-local attention technique have made significant contributions to the performance improvement of Bat-G2 net. This study clearly demonstrates the implementation feasibility of the new modality of 'seeing by hearing'.

ACKNOWLEDGMENT

(Seohyeon Kim and Gunpil Hwang are co-first authors.)

REFERENCES

[1] M. Kutila, P. Pyykonen, W. Ritter, O. Sawade, and B. Schaufele, "Automotive LIDAR sensor development scenarios for harsh weather conditions," in *Proc. IEEE 19th Int. Conf. Intell. Transp. Syst. (ITSC)*, Nov. 2016, pp. 265–270.

[2] M. B. Tahir and M. Abdullah, "Distance measuring (hurdle detection system) for safe environment in vehicles through ultrasonic rays," *Global J. Res. Eng.*, vol. 12, no. 1-B, pp. 14–21, Feb. 2012.
[3] G. Hwang, S. Kim, and H. M. Bae, "Bat-g net: Bat-inspired high-resolution 3D image reconstruction using ultrasonic echoes," in *Proc. Adv. Neural Infor. Process. Syst. (NIPS)*, 2019, pp. 3715–3726.
[4] J. A. Simmons, "Bats use a neuronally implemented computational acoustic model to form sonar images," *Current Opinion Neurobiol.*, vol. 22, no. 2, pp. 311–319, Apr. 2012.
[5] J. A. Simmons, M. J. Ferragamo, and M. I. Sanderson, "Echo delay versus spectral cues for temporal hyperacuity in the big brown bat, eptesicus fuscus," *J. Comparative Physiol. A, Sensory, Neural, Behav. Physiol.*, vol. 189, no. 9, pp. 693–702, Sep. 2003.
[6] G. Kaniak and H. Schweinzer, "A 3d airborne ultrasound sensor for high-precision location data estimation and conjunction," in *Proc. IEEE Instrum. Measur. Tech. Conf. (IMTC)*, May 2008, pp. 842–847.
[7] J. Steckel, A. Boen, and H. Peremans, "Broadband 3-D sonar system using a sparse array for indoor navigation," *IEEE Trans. Robot.*, vol. 29, no. 1, pp. 161–171, Feb. 2013.
[8] H. Akbarally and L. Kleeman, "A sonar sensor for accurate 3D target localisation and classification," in *Proc. IEEE Int. Conf. Robot. Autom.*, vol. 3, May 1995, pp. 3003–3008.
[9] A. Ochoa, J. Urena, A. Hernandez, M. Mazo, J. A. Jimenez, and M. C. Perez, "Ultrasonic multitransducer system for classification and 3-D location of reflectors based on PCA," *IEEE Trans. Instrum. Meas.*, vol. 58, no. 9, pp. 3031–3041, Sep. 2009.
[10] I. E. Dror, M. Zagaeski, and C. F. Moss, "Three-dimensional target recognition via sonar: A neural network model," *Neural Netw.*, vol. 8, no. 1, pp. 149–160, Jan. 1995.
[11] M. Moebus and A. Zoubir, "Three-dimensional ultrasound imaging in air for parking and pedestrian protection," in *In-Vehicle Corpus and Signal Processing for Driver Behavior*. New York, NY, USA: Springer, 2009, pp. 137–147.
[12] S. Watanabe and M. Yoneyama, "An ultrasonic visual sensor for three-dimensional object recognition using neural networks," *IEEE Trans. Robot. Autom.*, vol. 8, no. 2, pp. 240–249, Apr. 1992.
[13] P. Boufounos, "Compressive sensing for over-the-air ultrasound," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2011, pp. 5972–5975.
[14] P. A. Saillant, J. A. Simmons, S. P. Dear, and T. A. Mcmullen, "A computational model of echo processing and acoustic imaging in frequency-modulated echolocating bats: The spectrogram correlation and transformation receiver," *J. Acoust. Soc. Amer.*, vol. 94, no. 5, pp. 2691–2712, Nov. 1993.
[15] J. A. Simmons, "A view of the world through the bat's ear: The formation of acoustic images in echolocation," *Cognition*, vol. 33, nos. 1–2, pp. 155–199, Nov. 1989.
[16] F. Devaud, G. Hayward, and J. J. Soraghan, "The use of chirp overlapping properties for improved target resolution in an ultrasonic ranging system," in *Proc. IEEE Ultrason. Symp.*, vol. 3, Aug. 2004, pp. 2041–2044.
[17] G. Hayward, F. Devaud, and J. J. Soraghan, "P1G-3 evaluation of a bio-inspired range finding algorithm (BIRA)," in *Proc. IEEE Ultrason. Symp.*, Oct. 2006, pp. 1381–1384.
[18] G. Kaniak and H. Schweinzer, "A 3D airborne ultrasound sensor for high-precision location data estimation and conjunction," in *Proc. IEEE Instrum. Meas. Technol. Conf.*, May 2008, pp. 842–847.
[19] H. Peremans and J. Hallam, "The spectrogram correlation and transformation receiver, revisited," *J. Acoust. Soc. Amer.*, vol. 104, no. 2, pp. 1101–1110, Aug. 1998.
[20] J. A. Simmons, P. A. Saillant, J. M. Wotton, T. Haresign, M. J. Ferragamo, and C. F. Moss, "Composition of biosonar images for target recognition by echolocating bats," *Neural Netw.*, vol. 8, nos. 7–8, pp. 1239–1261, Jan. 1995.
[21] O. Heson, Jr., "Biosonar imaging of insects by Pteronotus p. parnelli, the mustached bat," *Natl. Geogr. Res*, vol. 3, pp. 82–101, 1987. [Online]. Available: https://scholar.google.com/scholar_lookup?journal=Natl.+Geog.+Res&title=Biosonar+imaging+of+insects+by+Pteronotus+parnellii,+the+mustached+bat&author=OW+Henson&author=A+Bishop&author=A+Keating&author=J+Kobler&volume=3&publication_year=1987&pages=82-101&
[22] R. Kober and H. Schnitzler, "Information in sonar echoes of fluttering insects available for echolocating bats," *J. Acoust. Soc. Amer.*, vol. 87, no. 2, pp. 882–896, Feb. 1990.
[23] L. A. Miller and S. B. Pedersen, "Echoes from insects processed using time delayed spectrometry (TDS)," in *Animal Sonar*. New York, NY, USA: Springer, 1988, pp. 803–807.
[24] H. U. Schnitzler, D. Menne, R. Kober, and K. Heblich, "The acoustical image of fluttering insects in echolocating bats," in *Neuroethology and Behavioral Physiology*. New York, NY, USA: Springer, 1983, pp. 235–250.
[25] J. A. Simmons and L. Chen, "The acoustic basis for target discrimination by FM echolocating bats," *J. Acoust. Soc. Amer.*, vol. 86, no. 4, pp. 1333–1350, Oct. 1989.
[26] J. A. Simmons, "The processing of sonar echoes by bats," in *Animal Sonar Systems*. New York, NY, USA: Springer, 1980, pp. 695–714.
[27] J. A. Simmons and R. A. Stein, "Acoustic imaging in bat sonar: Echolocation signals and the evolution of echolocation," *J. Comparative Physiol. A*, vol. 135, no. 1, pp. 61–84, 1980.
[28] G. Cybenko, "Approximation by superpositions of a sigmoidal function," *Math. Control, Signals, Syst.*, vol. 2, no. 4, pp. 303–314, Dec. 1989.
[29] J. Ba, V. Mnih, and K. Kavukcuoglu, "Multiple object recognition with visual attention," 2014, *arXiv:1412.7755*. [Online]. Available: http://arxiv.org/abs/1412.7755
[30] J. Fu, H. Zheng, and T. Mei, "Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4438–4446.
[31] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3156–3164.
[32] K. Hara, M.-Y. Liu, O. Tuzel, and A.-m. Farahmand, "Attentional network for visual object detection," 2017, *arXiv:1702.01478*. [Online]. Available: http://arxiv.org/abs/1702.01478
[33] J. Li, Y. Wei, X. Liang, J. Dong, T. Xu, J. Feng, and S. Yan, "Attentive contexts for object detection," *IEEE Trans. Multimedia*, vol. 19, no. 5, pp. 944–954, May 2017.
[34] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6077–6086.
[35] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, and T.-S. Chua, "SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5659–5667.
[36] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2015, pp. 2048–2057.
[37] M. I. Hasan Chowdhury, K. Nguyen, S. Sridharan, and C. Fookes, "Hierarchical relational attention for video question answering," in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2018, pp. 289–297.
[38] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.
[39] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7794–7803.
[40] E. Covey and J. H. Casseday, "The lower brainstem auditory pathways," in *Hearing by Bats*. New York, NY, USA: Springer, 1995, pp. 235–295.
[41] W. Shi, J. Caballero, F. Huszar, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1874–1883.
[42] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[43] M. E. Bates, S. A. Stamper, and J. A. Simmons, "Jamming avoidance response of big brown bats in target detection," *J. Experim. Biol.*, vol. 211, no. 1, pp. 106–113, Jan. 2008.
[44] N. Ulanovsky, M. B. Fenton, A. Tsoar, and C. Korine, "Dynamics of jamming avoidance in echolocating bats," *Proc. Roy. Soc. London B, Biol. Sci.*, vol. 271, no. 1547, pp. 1467–1475, Jul. 2004.

[45] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.
[46] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 3–19.

SEOHYEON KIM (Member, IEEE) received the B.S. degree in electronic and electrical engineering from Korea University, Seoul, South Korea, in 2013, and the M.S. and Ph.D. degrees in electrical engineering from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea, in 2015 and 2020, respectively. His research interests include bat-inspired high-resolution 3D ultrasound imaging systems (autonomous vehicle, drone, and robot), artificial intelligence (AI), artificial neural networks, design of energy-efficient architectures, implementation on integrated circuits for machine learning, and biomedical applications. He received the Bronze Award from the 22nd Samsung Humantech Paper Award in 2020 and the Outstanding Paper Award from the Qualcomm-KAIST Innovation Awards in 2019.

GUNPIL HWANG received the B.S., M.S., and Ph.D. degrees in electrical engineering from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea, in 2013, 2015, and 2020, respectively. He is currently a Postdoctoral Researcher with KAIST. His research interests include bat-inspired high-resolution 3D ultrasound imaging systems (autonomous vehicle, drone, and robot), artificial intelligence (AI), artificial neural networks, and biomedical circuits. He received the Bronze Award from the 22nd Samsung Humantech Paper Award in 2020, the Outstanding Paper Award from the Qualcomm-KAIST Innovation Awards in 2019, and the Global Ph.D. Fellowship from the National Research Foundation (NRF) from 2015 to 2017. He was a co-recipient of the Gold Prize from the 22nd Samsung Humantech Paper Award in 2015, the Outstanding Paper Award from the Qualcomm Innovation Awards in 2015, and the Bronze Award from the 15th Korea Intellectual Property Office (KIPO) Circuit Design Contest in 2014.

HYEON-MIN BAE (Member, IEEE) received the B.S. degree in electrical engineering from Seoul National University, Seoul, South Korea, in 1998, and the M.S. and Ph.D. degrees in electrical and computer engineering from the University of Illinois at Urbana-Champaign, Champaign, IL, USA, in 2001 and 2004, respectively. From 1995 to 1996, he served his Military duty in Dokdo. From 2001 to 2007, he led the analog and mixed-signal design aspects of OC-192 MLSE-based EDC ICs with Intersymbol Communications Inc., Champaign. From 2007 to 2009, he was with Finisar Corporation (NASDAQ: FNSR), Sunnyvale, CA, USA, after its acquisition of Intersymbol Communications Inc. Since 2009, he has been with the Faculty of the Electrical Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea, where he is currently a Professor. In 2010, he founded Terasquare Inc., Seoul, which provided low-power 100 Gb/s transceiver solutions. Teresquare Inc., was acquired by Gigpeak (NYSE:GIG), in 2015. In 2013, he founded OBElab Inc., Seoul, a bio start-up that manufactures portable functional brain-imaging systems. In 2016, he also founded Point2technology Inc., San Jose, CA, USA, that provides dielectric waveguide-based high-speed interconnect solutions. His research interests include wireline communication and medical systems. He was on the Wireline Subcommittee of the International Solid-State Circuit Conference (ISSCC) from 2014 to 2019. He received the Excellence Award from the National Academy of Engineering of Korea in 2013 and the 2006 IEEE JOURNAL OF SOLID-STATE CIRCUITS Best Paper Award. He served as a Distinguished Lecturer for the IEEE Solid-State Circuit Society from 2017 to 2019.