Bioinspiration & Biomimetics



# Small-scale location identification in natural environments with deeplearning based on biomimetic sonar echoes

To cite this article before publication: Liujun Zhang *et al* 2023 *Bioinspir. Biomim.* in press https://doi.org/10.1088/1748-3190/acb51f

# Small-scale location identification in natural environments with deep learning based on biomimetic sonar echoes

**Liujun Zhang[1], Andrew Farabow[2], Pradyumann Singhal[2], and Rolf Müller[3,*]**

[1]Department of Electrical and Computer Engineering, Virginia Tech, Blacksburg, VA 24060, USA
[2]Department of Computer Science, Virginia Tech, Blacksburg, VA 24060, USA
[3]Department of Mechanical Engineering, Virginia Tech, Blacksburg, VA 24060, USA
[*]rolf.mueller@vt.edu

## ABSTRACT

Many bat species navigate in complex, heavily vegetated habitats. To achieve this, the animal relies on a sensory basis that is very different from what is typically done in engineered systems that are designed for outdoor navigation. Whereas the engineered systems rely on data-heavy senses such as lidar, bats make do with echoes triggered by short, ultrasonic pulses. Prior work has shown that "clutter echoes" originating from vegetation can convey information on the environment they were recorded in – despite their unpredictable nature. The current work has investigated the spatial granularity that these clutter echoes can convey by applying deep-learning location identification to an echo data set that resulted from the dense spatial sampling of a forest environment. The GPS location corresponding to the echo collection events was clustered to break the survey area into the number of spatial patches ranging from two to 100. A convolutional neural network (Resnet 152) was used to identify the patch associated with echo sets ranging from one to ten echoes. The results demonstrate a spatial resolution that is comparable to the accuracy of recreation-grade GPS operating under foliage cover. This demonstrates that fine-grained location identification can be accomplished at very low data rates.

## Introduction

The ability to navigate autonomously in natural, vegetated environments could enable a multitude of technical applications that include search and rescue[1–3], precision agriculture[4,5], as well as environmental surveillance[6]. A fundamental ability that is needed for many of the navigation challenges posed by outdoor applications of autonomous systems is identifying a location, e.g., for the purpose of building a map of the environment where the system is supposed to operate.

The most common approach to identifying a given location is to utilize the Global Positioning System (GPS,[7]). However, GPS has its limitations: There are a number of environments where GPS is not available at all, e.g., under water[8,9] or in caves and mines[10,11]. In addition, GPS has been shown to be vulnerable to jamming or manipulation[12]. But even under conditions where an unadulterated GPS signal could be accessed in principle, there can be circumstances that result in a substantially reduced accuracy, e.g., under dense foliage cover[13], in indoor environments[14], or in urban canyons with dense foliage cover[15].

An alternative to location identification with GPS is provided by vision-based landmark recognition[16,17]. These approaches typically rely on the recognition of object shapes by virtue of matching deterministic templates. While it is easy to see how template matching would work for man-made landmarks such as buildings that have clearly recognizable shape patterns, plants in natural vegetation have complex shapes with a large amount of randomness[18] that may not be easily captured by a deterministic image template. Template matching based on three-dimensional optical recordings of an environment, e.g., using lidar[19,20], are likely to suffer from the same problem. In addition, navigation based on laser scans requires acquiring, storing, and processing large amounts of data. A single lidar sensor, for example, can produce data rates of about 23 $\frac{\text{Mbit}}{\text{s}}$[21,22]. The large computational cost and power consumption associated with handling such data rates may be difficult to satisfy in the context of small autonomous systems.

Echolocating bats are capable of navigation in a wide variety of natural habitats that include densely vegetated environments[23,24]. In addition, bats have been shown to travel distances as large as 50 km in a single night and then return to their roosts in the morning[25,26]. Hence, bats from these species must be able to create maps of their environments that allow them to find their ways to their feeding grounds and back to their roosts.

The ability of bats to navigate in densely vegetated environments based on biosonar is of particular interest because of the special nature of sonar echoes from natural vegetation[27,28]. A typical foliage is composed of a multitude of leaves and other

sound-reflecting elements. Each of these elements contributes to the received echoes based on its position, orientation, and shape. Since all the specific values of the parameters that the determine the reflection from an individual leaf remain unknown, a foliage is best approximated as a stochastic array of reflectors[27,29] that results in likewise unpredictable echo waveforms[27]. For sonar-based navigation, the implication of this randomness is that any sonar system will never see the same echo waveform again[30]. Hence, conventional template-based methods for recognition of a location-specific pattern will not work. Despite these difficulties, the biosonar abilities of bats demonstrate that sonar-based solutions to the location-finding problem must exist and can be realized in a highly reliable and parsimonious fashion.

As a first step towards replicating the biosonar-based location-finding skills of bats, prior work by the authors[31] has shown that different habitats identification of different locations in natural environments based on single (15 ms) echoes is possible using deep learning methods based on time-frequency representations of the echoes. In this study, it was possible to not only identify ten different locations that were spaced within a 50-kilometer diameter, but also neighboring walking trails at the same habitat. The latter results have hinted at the possibility that biomimetic sonar echoes can convey location information with much finer resolution.

To examine the spatial resolution for locations that biomimetic sonar could provide in natural structure-rich habitats, the current work has characterized biomimetic echo data that collected to cover a contiguous natural forest area with a dense set of echo measurements. The echoes of in the previous study were labelled based on different collection locations that were separated by distances of several kilometers (in case of the different sites) or at least several ten meters (in case of different walking trains). Hence, these data sets are likely to have captured large-scale differences between entirely different vegetation types or at least distinct local variants of the same vegetation type (e.g., due to different soil or exposure).

Hence, the goal of the current study has been to examine the small-scale granularity of the location information that can be extracted from echoes collected within a contiguous area that was covered by the same vegetation type. To this end, each of the acquired echoes was labelled with geographic coordinates provided by a GPS. This GPS data was then used to cluster the locations into small patches and determine the level of spatial granularity can be resolved based on deep-learning classification of the echoes.

## Materials and Methods

### Biomimetic sonar

A biomimetic sonar head (Fig. 1A,[31]) was used to collect all foliage echoes for the present study. The system consisted of a sonar emitter and two receivers, a top-level controller, a set of microcontrollers used for digital-to-analog and analog-to-digital conversion, cameras to record images for documentation purposes, a GPS, as well as a power system to support all these devices.

For the sonar system proper, the main components were two electrostatic ultrasonic loudspeakers (diameter 3.8 cm, 600 Series, SensComp Inc., Livonia, MI, USA) with a peak response frequency located at ∼50 kHz and a -6 dB passband extending from ∼40 up to ∼80 kHz. Two capacitive MEMS microphones (Monomic, Dodotronic, Rome, Italy, -6 dB frequency range from ∼2 to ∼125 kHz) were used to receive the returning echoes.

The top-level computer of the sonar head (Raspberry Pi 3, Model B+, RS Components, Cambridge, UK) was used to control data collection via a user interface, store digital data from the microphones, and to visualize the incoming echoes in approximate real-time. A microcontroller (Arduino Due, Arduino, Somerville, MA, USA, clock frequency 84 MHz) was tasked with handling digital-to-analog and analog-to-digital conversion of the sonar pulses and the returning echoes respectively. The emitted pulses were converted to analog input signals for the ultrasonic transducer with a sampling rate of 1.6 MHz (the device maximum) and 12 bits amplitude resolution. The conversion of the microphone outputs to digital signal representations were conducted with a sampling rate of 400 kHz per channel and with an amplitude resolution of 16 bits.

The entire system was powered by a DC battery (Lithium Polymer RC Battery, 22.2 V, 4.5 Ah, Floureon, Nantou, Taiwan). A GPS module (Adafruit Ultimate GPS, Breakout 3, Adafruit Industries, New York, NY, USA) recorded the geographic coordinates associated with each echo collection site. The GPS had a manufacturer-specified position error of 1.8 m[32]. To assess whether specification was valid for the recording conditions of the experiments reported here, 100 GPS data points were recorded along a straight paved road that was aligned with the edge of the vegetation in which the echoes were recorded. Under the assumption that the road edge can be described by a straight line, the root-mean-square error associated with fitting a line to this position data can be used as a measure for the accuracy of the GPS. This error was found to be about 5.6 m.

Finally, all data acquisition work was documented with a stereo-pair of two video cameras (HERO 3, GoPro, San Mateo, CA, USA) that were mounted on the biomimetic sonar head, faced in the same direction as the transducers, and recorded videos during echo data collection. None of the recorded videos were subjected to any form of quantitative analysis in this study.

## Data collection

The echo data was collected in a natural wooded area (known as the "Stadium Woods", Fig. 1D) on the Virginia Tech campus in Blacksburg, Virginia. The size of the study site was approximately 180 by 150 m. In general, the terrain of the study site was slightly rolling with a uniform vegetation cover of mature forest and substantial amounts of undergrowth (Fig. 1B). However, some small spots could not be walked safely due to the presence of local obstacles such as boulders or ravines and were hence left out of the data acquisition. An estimate of the area covered by the echo recordings has been derived from the measured GPS positions using a morphological closing operation which consists of a dilation followed by an erosion to determine a contiguous area covered by points[33]. The input data for this operation were the geographical positions associated with the echo recordings as determined by a reading from the GPS receiver. For the closing operation, each GPS location was represented by a point corresponding to a radius of 1.5 m in the real world. The structuring element of the morphological closing operation, i.e., the mask that is used to probe the image, was a circle with a radius that corresponded to 2.5 m.

During data collection, the biomimetic sonar was hand-carried (to approximate the natural variability in the flight paths that bats might take through the forest) at a distance of about 1 to 1.5 m from the nearest vegetation. While being moved through vegetation, the biomimetic sonar head was also rotated by hand in scanning motions that covered azimuth as well as elevation. As for the walking path, these motions were intended to approximate how a bat's biosonar might scan its surroundings. The sampling paths were traversed with a constant walking speed of about 0.2 m/s while the data collection rate was about three echoes per second. The GPS module was used to record the location of the sonar (Fig. 1C) with an update rate of 0.2 Hz. A second-order polynomial fit was used to interpolate the GPS data for each instant of echo collection. The interpolated GPS locations were then attached to each recorded echo as the label for supervised learning.

The pulse waveform that was used to trigger the vegetation echoes was inspired by the biosonar pulses of constant frequency (CF) - frequency-modulated (FM) bats that combine narrow-band and frequency-modulated portions[34]. To mimic both of these signal components, the first 7 ms portion of the emitted pulses consisted of an FM chirp that sweeped from 55 kHz down to 45 kHz (Fig. 2, black solid box). The second part consisted of an CF signal centered at 60 kHz with a duration of 5 ms (Fig. 2, white solid box). Hence, the FM and CF components of the signals were hence not connected in frequency which does match the continuity in the time-frequency contours of bat biosonar pulses, but made separating the CF and FM components easier. The total length of each echo recording was 25 ms. Since each echo recording was started with the beginning of the respective pulse, it including the two direct transmissions from the speaker as well as the reflected echoes. The resulting echoes (Fig. 2, dashed box) were used as input for data classification.

## Clustering of the GPS data

The study area was broken up into a set of coherent spatial patches based on the GPS coordinate data using a clustering approach (MiniBatch k-means,[35], implemented in the sklearn Python library[36]). The MiniBatch approach reduces the computation time from that of k-means by processing random batches of data with a fixed size small enough so that they can be stored in memory. Clustering is repeated in an iterative process where each iteration is based on a random pick of samples and the iteration stops once the convergence criterion is reached. Like in regular k-means, the objective of the algorithm is to minimize the within-cluster sum of squares. For each clustering attempt, the centroids of the clusters were randomly initialized three times, then the algorithm picked the best of the initialization as measured by the sum-of-square distances to their cluster center. To prevent premature stopping, the maximum consecutive number of mini-batches that do not yield an improvement on the figure of merit was set to 10. The size of the mini-batches was set to 256 sample points, and the maximum number of iterations to 100.

The silhouette value[37] was used as a measure of how coherent the generated spatial clusterings were. It measures how similar the elements within a cluster are to each other (cohesion) compared to how similar elements are across different clusters (separation). The silhouette value was calculated by the difference between the average within-cluster and cross-cluster distances normalized by the maximum value of these distances. Hence, the silhouette value ranges from $-1$ to $+1$, where a high value indicates that the elements are well matched to their respective clusters and poorly matched to neighboring clusters(Fig. 3).

Using this approach, the GPS data were clustered into different numbers of coherent spatial "patches" that ranged from a minimum of two up to a maximum of 100 in number (Fig. 4). For each desired number of patches, the clustering was repeated 100 times, and the result with the highest silhouette value was retained for the subsequent analysis. In addition to the silhouette value, the distribution of sample (Fig. 5) locations across the different clusters was monitored to ensure that it was approximately even (Fig. 6).

## Acoustical signal processing

In the first step of processing the echoes, a bandpass filter (finite impulse response, FIR, filter design based on a 256-point Hamming window with 50% overlap) was used to extract the frequency band occupied by the employed pulses from the echo recordings. For the FM pulses, this passband ranged from 40 to 58 kHz (-3 dB corner frequencies). The same bandpass filter design was used for the CF pulses, but in this case, the -3 dB passband covered the frequency range from 58 to 62 kHz.

The bandpass-filtered echoes were converted into spectrogram representations (Hanning window with a length of 256 samples, FFT length 256 samples, 50% window overlap). The spectrogram representations were cropped along the frequency axes to the passband of the respective pulses. Hence, for the FM pulses (frequency range from 45 to 55 kHz), the spectrogram matrix size was $18 \times 15$ whereas for the CF pulses (58 to 62 kHz), it was $7 \times 11$. For each time-frequency bin, i.e., pixel, in the spectrogram representation, the respective power spectral density was represented by an eight-byte floating-point number. The cropped spectrogram images served as input for classifying the echoes into the corresponding spatial patches.

## Deep-learning for location classification

The network used for patch classification was inspired by a state-of-the-art convolutional deep neural network for image classification (ResNet152,[38]). For the current work, the published ResNet152 architecture was modified by reducing the initial kernel size from 7×7 to 3×3 pixels, removing the stride, and adjusting the pooling layers size from 3×3 to 2×2 pixels. The deep neural network was implemented in TensorFlow (version 1.13.2, Google Brain Team,[39]) via the Keras interface library (version 2.3.1, F. Chollet,[40]) and the Python programming language (version 3.7).

The deep neural network used (Fig. 7) started with the full input spectrogram size (18×15 pixels for FM pulses, 7×11 pixels for CF pulses), followed by an initial convolution layer, batch normalization[41], an activation function (rectified linear unit, ReLu,[42]), and a single maximum-pooling layer (2×2 pixels for FM only,[43]). These initial layers were followed by 50 serial repetitions of a basic unit that contained a parallel arrangement of an identity block and a convolution block. The identity block insured that the original image features were also passed along into the network. The outputs of the identity and convolution blocks were concatenated as they were passed on from unit to unit. The convolution block of each unit consisted of three convolutional layers, each followed by batch normalization and a ReLu. The final layers of the network consisted of an average pooling layer followed by a 1000 nodes fully connected (fc1000) layer and finally an output prediction layer based on the Softmax function[44] that produced the estimated numbers of the spatial patch the respective echo originated from.

Since bats produce pulse trains with repetition rates up to at least 14 pulses per second [45–47]and hence should have ready access to multiple echoes to determine their location. To mimic this, deep-learning classification has been attempted on sets of echoes that contained 2 to 10 echoes subsequently recorded in the same spatial patch. To this end, a deep neural network has been designed for integrating multiple inputs (Fig. 8. The first part of this network has been based on ResNet152 and served to extract the features from the spectrograms as was done for classification based on single echoes. To enable the processing of multiple echoes, the outputs from the SoftMax layer of the initial ResNet152 for each of the echoes were concatenated for all echoes processed in a given batch (i.e., 2 to 10 echoes) into a single vector. These concatenated vectors were then fed into a multilayer perceptron (MLP, [48]), which performed the classification of the spatial patch based on this aggregate vector. The MLP contained three layers, the first layer had a dimension of 512 nodes, the second had 256 nodes, and the final layer was the SoftMax layer with one node for each spatial patch. The maximum across the outputs from the SoftMax layer was used to estimate the spatial patch that a given set of echoes originated from.

For performing the deep-learning experiments, the entire echo data set was randomly partitioned so that 85% of the recorded echoes were used for training and the remaining 15% were used for testing. During the entire process, a five-fold cross-validation[49] was used for characterizing the classification performance of the respective network. Cross-entropy loss[50] was used as the loss function for training the deep neural networks. This loss was monitored along prediction accuracy during the training process to establish convergence and check for overfitting (Fig. 9).

Confusion matrices showing the distribution of pairs of estimated and actual patch numbers were used to compare the results across different numbers of spatial patches. In these experiments, the number of spatial patches ranged from 2 to 100 and the number of echoes used to classify any given patch from one to 10. (Fig. 10).

## Classification visualization

Saliency maps[51] were used to visualize which parts of the spectrogram were most important to the classifier networks in making a decision on the spatial patch a given echo or set of echoes came from. These maps are derived from the gradients of the network output over its input[51]. Here, the saliency values were computed from the final convolution layer that preceded the computation of the SoftMax weights. An average saliency map was compiled from the maps obtained across the different spatial patches (Fig. 11). The spectrogram input was from the pure echo part (Fig. 2, dashed black box), and the average was taken over all 2,000 echoes in each patch. For each echo, the absolute value of the weight matrix in the final convolution layer was used to calculate the average saliency map and the average saliency map was normalized to compare the differences from each patch.

In order to test whether the time-frequency bins with high saliency values contained indeed more information that was relevant to classification of the spatial patches than those bins with low saliency values, the spectrograms were portioned into bins associated with high and low saliency values. In particular, the time-frequency bins associated with the top 50% and the bottom 50% saliency values for the classification of six different patches were selected from the spectrograms (Fig. 12). Since

the saliency maps were calculated separately for the echo set from each patch, they were combined here using an intersection, i.e., the set of time-frequency bins where all individual maps had an above 50 % threshold saliency value. A multi-layer perceptron was trained in a supervised learning paradigm to estimate the spatial patches associated with each echo based on the spectrogram values in the time-frequency bins associated with either high or low saliency values. The utilized MLP architecture had the following layers: first, the spectrogram amplitudes in the selected time-frequency region were flattened a vector. Second, upsampling the time-frequency bins to 256 nodes. Then followed by a downsampling layer with 128 nodes. The last layer has the same node number with the number patches with the SoftMax as the activation function.

## Results

During the foliage scans, a total of 37,136 echo recordings were collected from each microphone channel along with 2,280 GPS unique locations. The area covered by this data collection was estimated to cover approximately 13,400 $m^2$, based on the morphological closing method.

The GPS locations could be clustered into coherent spatial patches. The uniformity in the size of these patches in terms of the number of GPS locations they contained depended on the number of patches with configurations containing fewer patches being more uniform with a smaller number of patches resulting in a greater uniformity. Hence, the ratio between the maximum and the minimum number of locations per patch tended to increase with the number of patches used to divide the experimental area. For example, when classifying GPS data into 2 patches, this ratio was about $1.1\pm0.15$ with 100 repetitions, and when the number of patches was increased to 100, and the average ratio increased to $5.8\pm2.7$ with 100 repetitions (Fig. 5). It was possible to obtain clustering results with spatial patches that were uniform in size as well coherent by repeating each clustering 100 times and picking the minimum ratio between the largest and smallest cluster while at the same time ensuring that the silhouette value did not go below a threshold value of 0.4 that was set empirically to achieve this goal(Fig. 3).

The training of the ResNet152 converged after about 50 epochs. The loss decreased very sharply in the first 20 epochs and then started to level out and slowly converge with additional epochs. The training was stopped at 100 epochs due to a lack of clear improvements beyond this point for both training and validating data. Over-fitting was not evident in the training results since the accuracy difference between training and testing data remained less than 3% after 50 epochs (Fig. 9).

The classification performance for the deep neural network that was set up for processing multiple echoes depended on both the number of patches and on the number of echoes and in a systematic manner (Fig. 10): Providing the network with larger numbers of echoes resulted in a monotonic increase in prediction accuracy over the range of echo set sizes tested. However, this effect showed signs of saturation, especially for small numbers of patches (Fig. 10b). Similarly, the classification performance showed a monotonic decrease as the number of spatial patches to be classified increased (Fig. 10c). For the smallest number of patches (two) and a single-echo input, the network achieved 94.6% accuracy. When more echoes were added to the input, the performance quickly approached 100% accuracy (Fig. 10b). For the largest number of spatial patches tested (100), classification accuracy increased from 44% for a single echo to 83% accuracy for ten echoes.

The averaged saliency maps obtained indicated that the regions near the beginning (i.e., within the first 3 milliseconds of the echo) where the most important to the classifiers (Fig. 11). Beyond this commonality, the maps also showed patterns that were at least somewhat specific to the respective patch (Fig. 11). These patch-specific differences were found to be in the location of the most salient regions along the frequency axis as well as in the different spread of these regions in time as well as frequency.

The results from the classification experiments based on the parts of the spectrogram that had either particularly or particularly low saliency values (Fig. 12) confirmed that the saliency maps did capture some of the distribution of classification-relevant information in the joint time-frequency domain. For example, using a three-layer MLP, five patches, and a single echo input resulted in 91% classification accuracy when the power-spectral density in the intersection of the top 50% saliency values were used. By comparison, the Resnet152-based classifier that was trained on the entire spectrogram matrix as its input performed only about 2% better than the three-layer MLP while taking much longer to train. For the intersection of the bottom 50% saliency values across the five patches, the accuracy of the MLP was just 62%.

## Results

During the foliage scans, a total of 37,136 echo recordings were collected from each microphone channel along with 2,280 GPS unique locations. The area covered by this data collection was estimated to cover approximately 13,400 $m^2$, based on the morphological closing method.

The GPS locations could be clustered into coherent spatial patches. The uniformity in the size of these patches in terms of the number of GPS locations they contained depended on the number of patches with configurations containing fewer patches being more uniform with a smaller number of patches resulting in a greater uniformity. Hence, the ratio between the maximum and the minimum number of locations per patch tended to increase with the number of patches used to divide the experimental area. For example, when classifying GPS data into 2 patches, this ratio was about $1.1\pm0.15$ with 100 repetitions, and when

the number of patches was increased to 100, and the average ratio increased to 5.8±2.7 with 100 repetitions (Fig. 5). It was possible to obtain clustering results with spatial patches that were uniform in size as well coherent by repeating each clustering 100 times and picking the minimum ratio between the largest and smallest cluster while at the same time ensuring that the silhouette value did not go below a threshold value of 0.4 that was set empirically to achieve this goal(Fig. 3).

The training of the ResNet152 converged after about 50 epochs. The loss decreased very sharply in the first 20 epochs and then started to level out and slowly converge with additional epochs. The training was stopped at 100 epochs due to a lack of clear improvements beyond this point for both training and validating data. Over-fitting was not evident in the training results since the accuracy difference between training and testing data remained less than 3% after 50 epochs (Fig. 9).

The classification performance for the deep neural network that was set up for processing multiple echoes depended on both the number of patches and on the number of echoes and in a systematic manner (Fig. 10): Providing the network with larger numbers of echoes resulted in a monotonic increase in prediction accuracy over the range of echo set sizes tested. However, this effect showed signs of saturation, especially for small numbers of patches (Fig. 10b). Similarly, the classification performance showed a monotonic decrease as the number of spatial patches to be classified increased (Fig. 10c). For the smallest number of patches (two) and a single-echo input, the network achieved 94.6% accuracy. When more echoes were added to the input, the performance quickly approached 100% accuracy (Fig. 10b). For the largest number of spatial patches tested (100), classification accuracy increased from 44% for a single echo to 83% accuracy for ten echoes.

The averaged saliency maps obtained indicated that the regions near the beginning (i.e., within the first 3 milliseconds of the echo) where the most important to the classifiers (Fig. 11). Beyond this commonality, the maps also showed patterns that were at least somewhat specific to the respective patch (Fig. 11). These patch-specific differences were found to be in the location of the most salient regions along the frequency axis as well as in the different spread of these regions in time as well as frequency.

The results from the classification experiments based on the parts of the spectrogram that had either particularly or particularly low saliency values (Fig. 12) confirmed that the saliency maps did capture some of the distribution of classification-relevant information in the joint time-frequency domain. For example, using a three-layer MLP, five patches, and a single echo input resulted in 91% classification accuracy when the power-spectral density in the intersection of the top 50% saliency values were used. By comparison, the Resnet152-based classifier that was trained on the entire spectrogram matrix as its input performed only about 2% better than the three-layer MLP while taking much longer to train. For the intersection of the bottom 50% saliency values across the five patches, the accuracy of the MLP was just 62%.

## Discussion

Prior work has already established that "clutter echoes" from natural vegetation contain information about the targets that produce them. This is true despite the profoundly unpredictable and unrepeatable nature of the individual waveforms[27]. Early on, it was demonstrated that different tree species can be distinguished based on echoes of their foliages[27,52]. Furthermore, prior work by the authors has already demonstrated that ten different locations in natural environments distributed over an area with a diameter of about 50 kilometers could be recognized reliably based on single clutter echoes[31]. It is possible that the differences in the echoes obtained across at least some of these sites were due to very dissimilar vegetation such as deciduous versus pine forest. In these cases, it can be expected that foliage types with large differences in parameters such as leaf size and density also give rise to very different echo waveforms. However, the same study has also demonstrated identification of two different tracks that were walked for echo collection at each of the ten sites[31]. The latter finding could be seen as an indication that location identification based on clutter echoes could be possible on a finer scale than would be defined based on fundamentally different vegetation types. Still, the different tracks of the previous study could reflect somewhat large-scale changes in the vegetation, e.g., due to differences in soil, exposure, or the level of maturity of a forest.

The ability to recognize locations on a large scale, e.g., by virtue of different vegetation types, is likely not enough to support an efficient navigation which should be able to chart a path to the destination in a continuous fashion and hence requires frequent, accurate, and precise updates on location. In this context, the results of the current study are significant because they demonstrate a much finer resolution that could very well support efficient navigation and hence explain how bats can find their way through the forest. They could also offer an interesting solution to the problem of navigation in GPS-denied environments[14,15] for man-made systems.

In the latter context, it is interesting to note that the spatial resolution achieved by the location patches that could be correctly identified here is not far from what has been reported for GPS operating under foliage cover: An evaluation of a recreation-grade GPS device (Suunto Ambit 3 Peak device) operated under foliage cover[13] has yielded RMS errors for location of 10.06 m for coniferous forest and 15.81 m for deciduous forest[53,54]. If the total area covered in the current work is broken up into 100 equal-area spatial patches, each patch would have a radius error of about 6 m. For this scenario, an accuracy higher than 85% was achieved based on sets of ten echoes. While it is difficult to do an exact comparison of these numbers and some of the scatter in the results presented here may actually go back to errors in the GPS reference, it appears that the biomimetic sonar-based localization explored here could achieve a similar accuracy than GPS.

Besides the spatial resolution for different locations, it is also worth considering the effort that is required for dealing with the data from different sensory modalities. State-of-the-art lidar systems, for example, can generate data rates between 20 and 100 Mbit/s (for systems with one to five sensors,[22]). Even higher data rates of 500 to 3,500 Mbits been reported for arrays of six to 12 cameras[22]. By comparison, each of the echo waveforms that were analyzed here just required 64 kbit of data (at 10 ms duration, 400 kHz sampling rate, and 16 bits resolution) to be represented without any compression. If the ten echoes that were used in the largest classification data sets in the present work were to be collected within one second, this would result in a data rate of 640 kbit/s. This would be just less than one thirtieth of the 20 Mbit/s data rate generated by a single lidar sensor. Hence, bioinspired approaches like the one explored here could offer much more parsimonious ways to support navigation than data-intensive sensors such as lidar. Operating on such low data rates could enable small, agile platforms that consume little power and are capable of fast reactions. In addition to the Low-SWaP (Low Size, Weight and Power), work on a different navigation problem, passageway finding[55], has found that the computationally expensive deep-learning methods could be replaced by a simple neuromorphic approach. This approach could be implemented in analog hardware and would hence be extremely fast and power-efficient. For the current task, location identification, the feasibility of such an approach has yet to be established. However, the localized nature of the relevant information in the time-frequency plane that was evident in the saliency maps could be conducive to encoding information that exists in a certain frequency channel and in a certain time interval using a simple spike code.

If bats are able to exploit the location information contained in the clutter echoes, it would provide an explanation for how the animals are able to find their way in densely vegetated environments without the need of reconstructing any deterministic features in their surroundings. Given the small size of brains in bats is about $0.82 \pm 0.21$ g[56] demonstrating this skill in bats would also be a strong indication that parsimonious implementations of the location estimates on clutter echoes are possible.

Future work on the ability to identify locations from clutter echoes should investigate the use of a better reference than consumer-grade GPS to better reference to evaluate resolution and accuracy that can be achieved. Ideally, such a detailed study of the resolution of the approach should be repeated across different habitats to see if some of them are better suited than others. An additional aspect that should investigated is the stability of location information over time. Since the informative echo properties do not rely on any deterministic spatial pattern, it could be hypothesized that they are very robust against changes to the positions of individual reflecting facets (e.g., leaves). However, seasonal changes to a foliage could disrupt identification. This is very obvious when considering a deciduous forest in summer and in winter, but even much more gradual changes may eventually degrade location identification. Finally, if location identification based on biomimetic echo is found to be sufficiently accurate and stable, it could be investigated how the specific nature of these echoes could be best integrated into state-of-the-art map-building approaches such as SLAM[57].

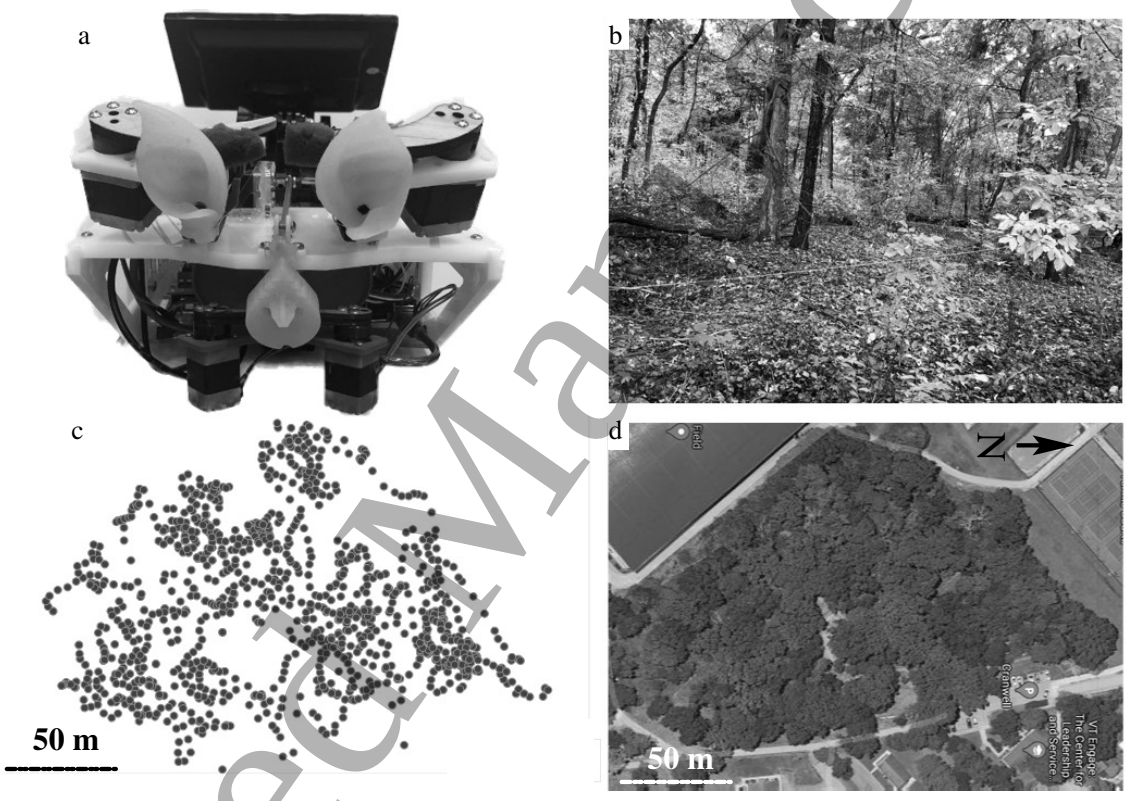


**Figure 1. Biomimetic sonar robot and field site used for data collection. (a)** Front view of the biomimetic sonar robot consisting of two ultrasonic loudspeakers to produce the pulses and two microphones mounted into the ears for echo reception, the screen in the back of the device provides the user interface. **(b)** Forest habitat at the field site. **(c)** GPS locations associated with the collected echo data set. **(d)** Satellite image of the entire data collection field site (size: 150 m by 180 m).

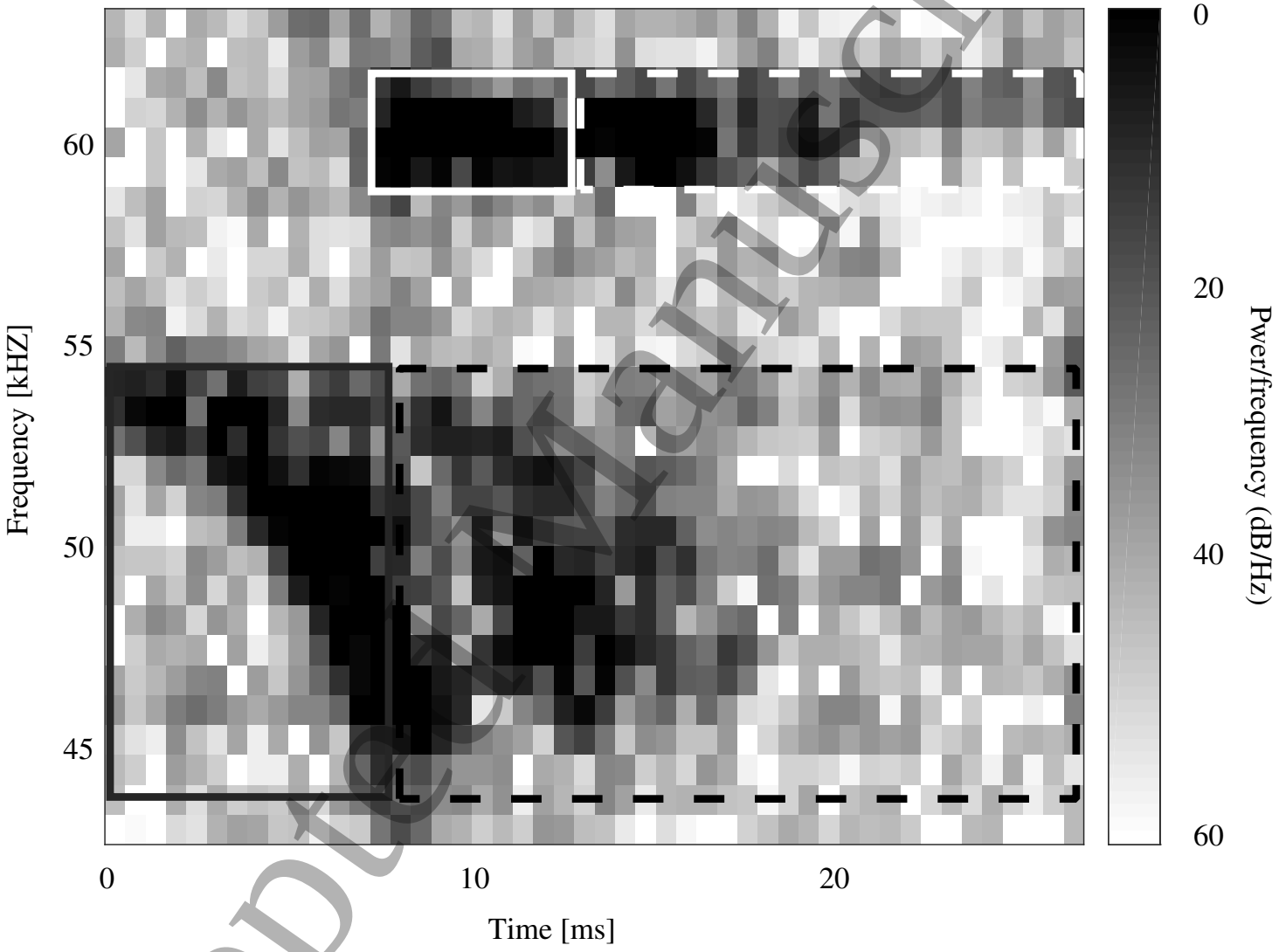


**Figure 2. Spectrogram of an example of the echoes that have been used for location identification.** The emitted signals consisted of a CF-FM pulse pairs where the FM pulse swept from 55 kHz down to 45 kHz over a duration of 7 ms (solid black box) and was followed by a CF pulse centered at 60 kHz and a duration of 5 ms (solid white box). The echoes to both pulses are shown in the dashed boxes (black dashes: FM echoes, white dashes: CF echoes).

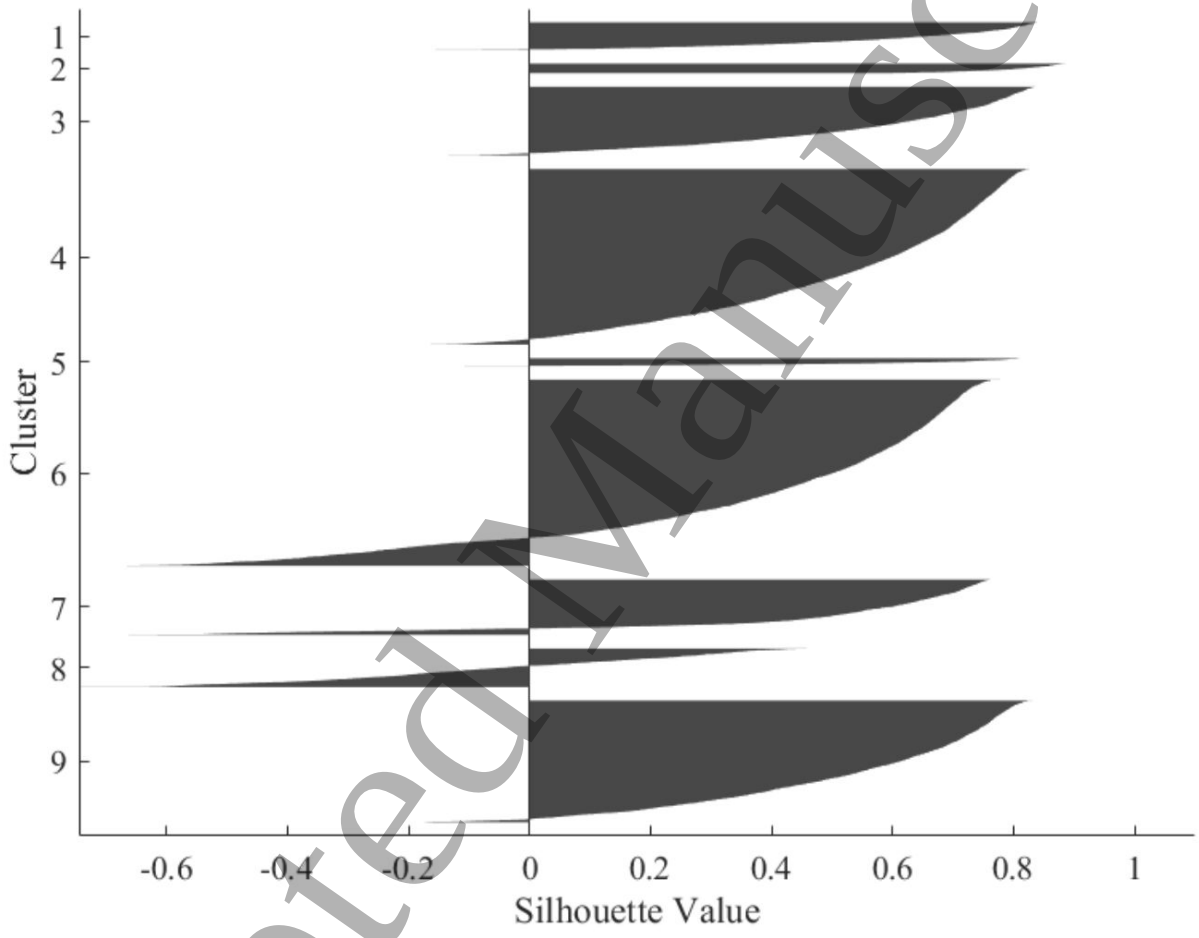


**Figure 3. Silhouette value distribution for nine clusters**. For each of the clusters, the silhouette values are shown for all the points within the respective cluster.

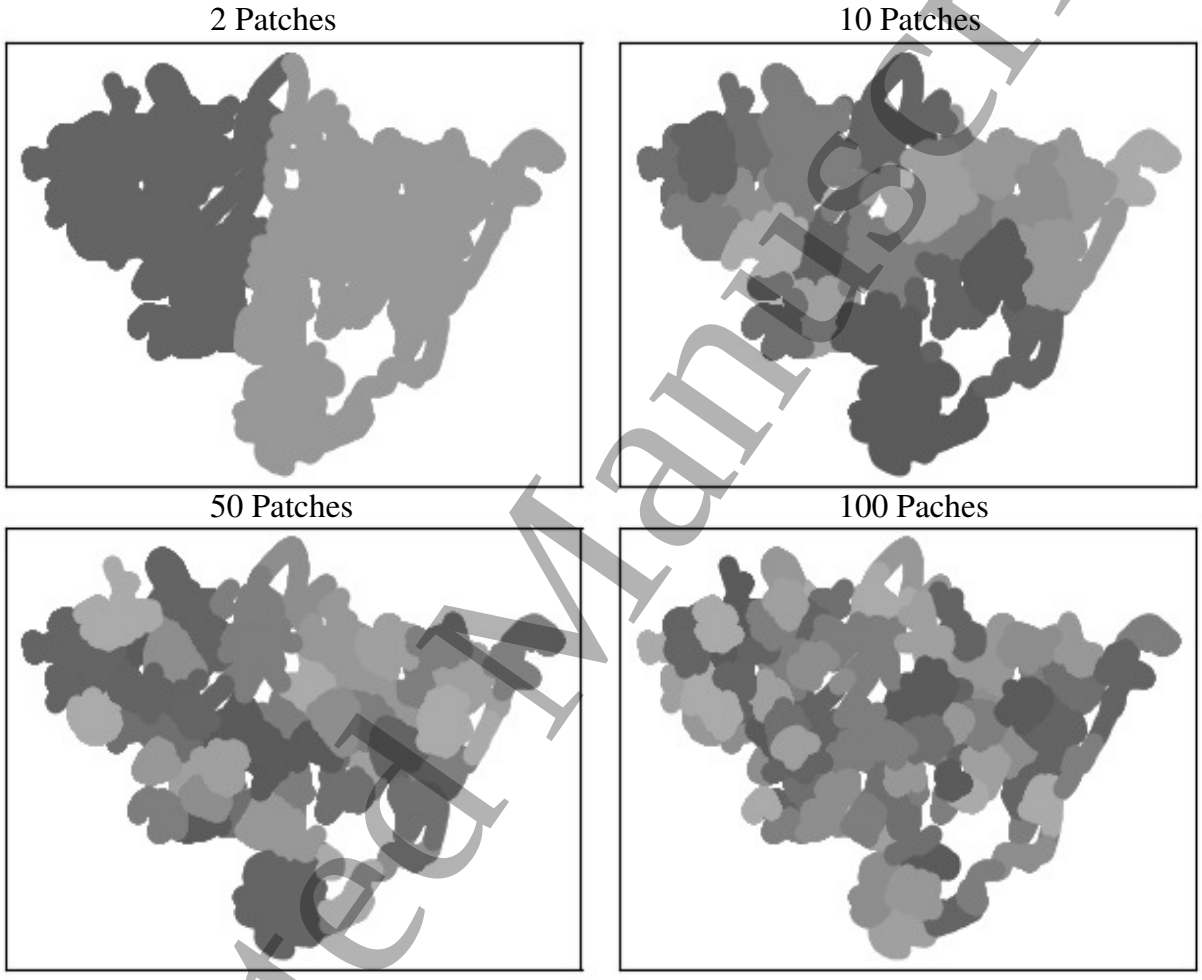


**Figure 4. Map of the GPS location data clustered into different numbers of spatial patches using the MiniBatch k-means method**. The clustering examples shown are for 2, 10, 50, and 100 distinct spatial patches. Different patches are marked by different gray levels).

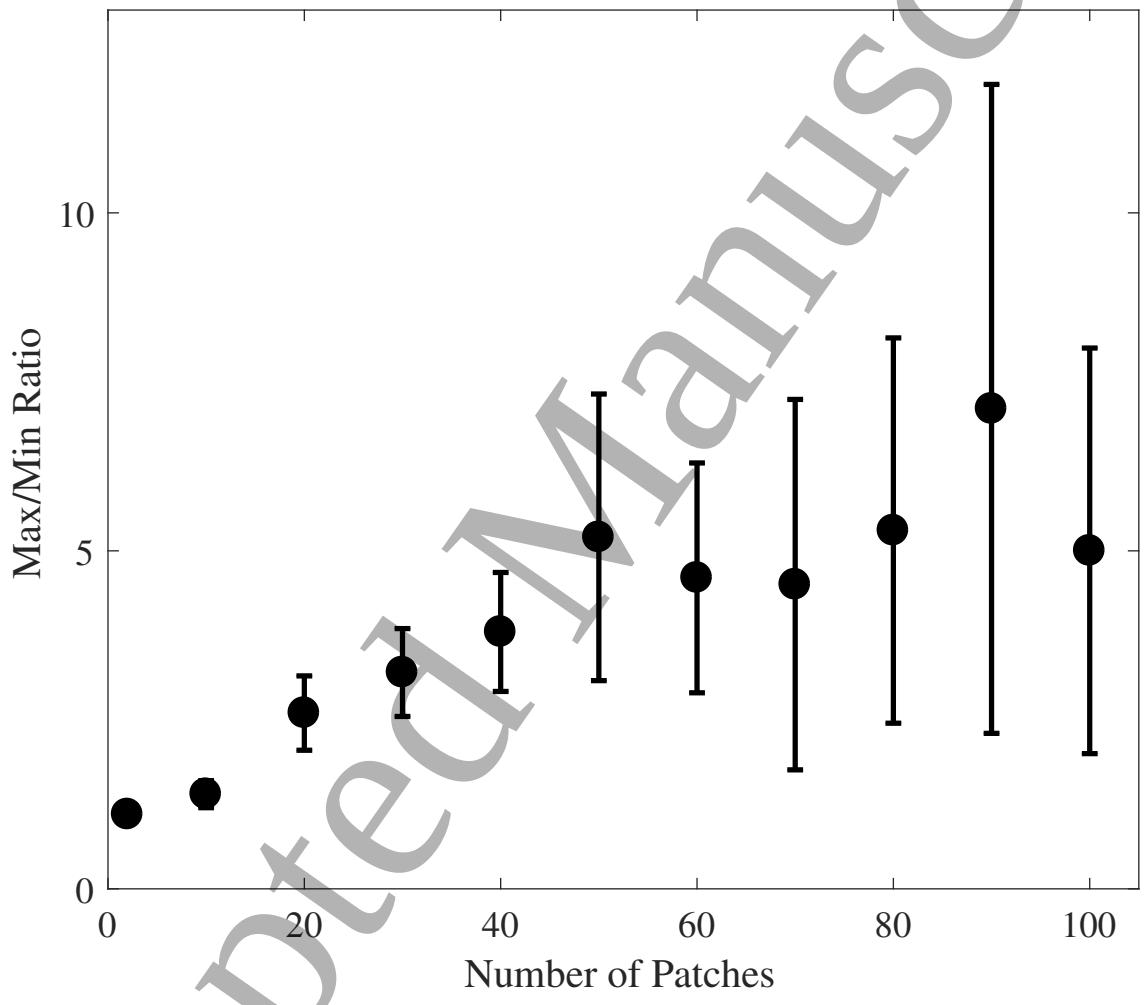


**Figure 5. Ratio of the maximum to maximum number of locations per cluster for different number of patches (*N*=100 repetitions).**

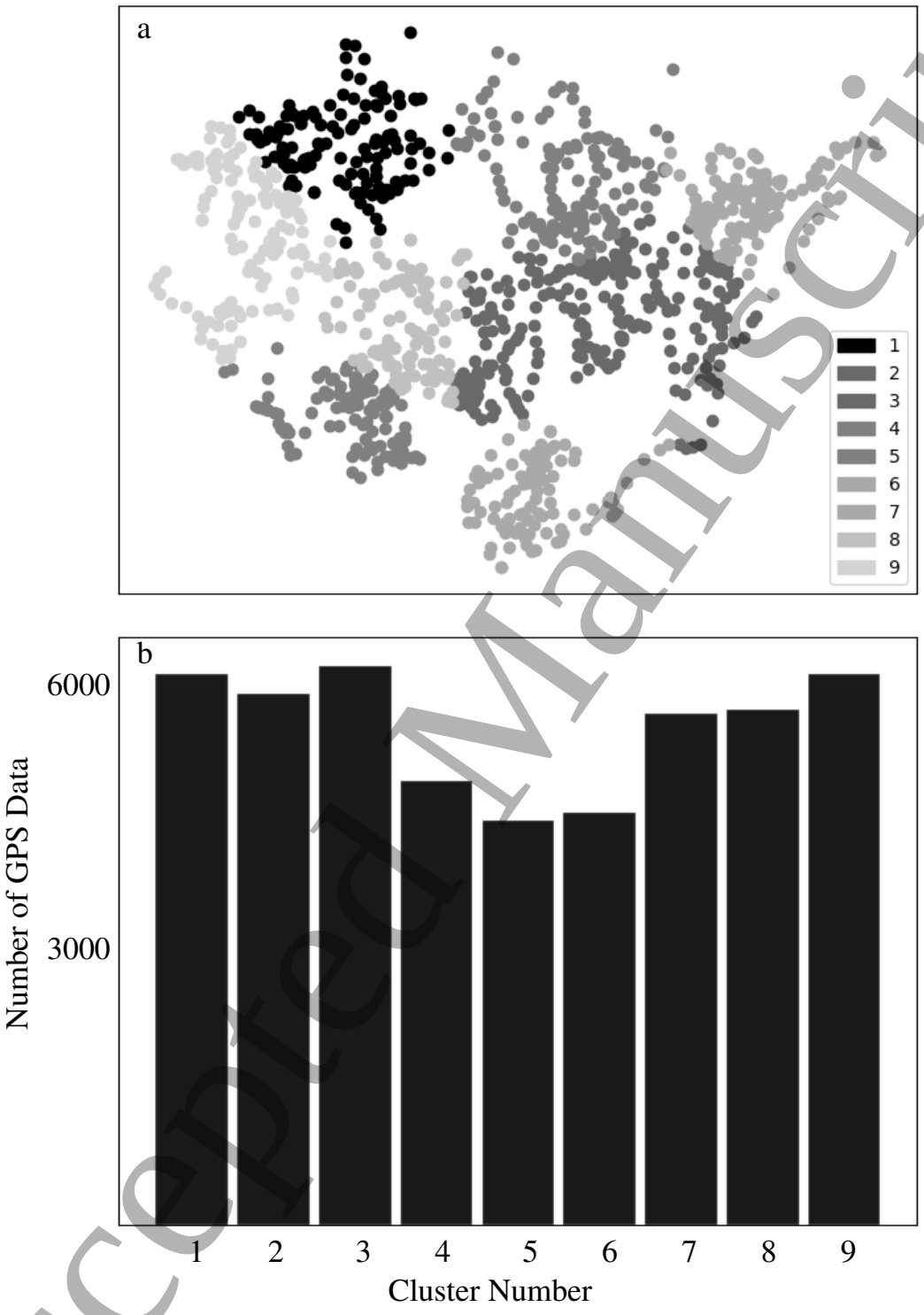


**Figure 6. Clustering the GPS locations into spatial patches while avoiding heavily skewed allocations across clusters.** **(a)** Map of the allocation of the GPS locations to different clusters (nine in this example, each marked by a different gray level). **(b)** Number of the GPS locations included in each cluster (spatial patch) with a maximum-to-minimum ratio for the number of locations per cluster of 1.38 in this example.

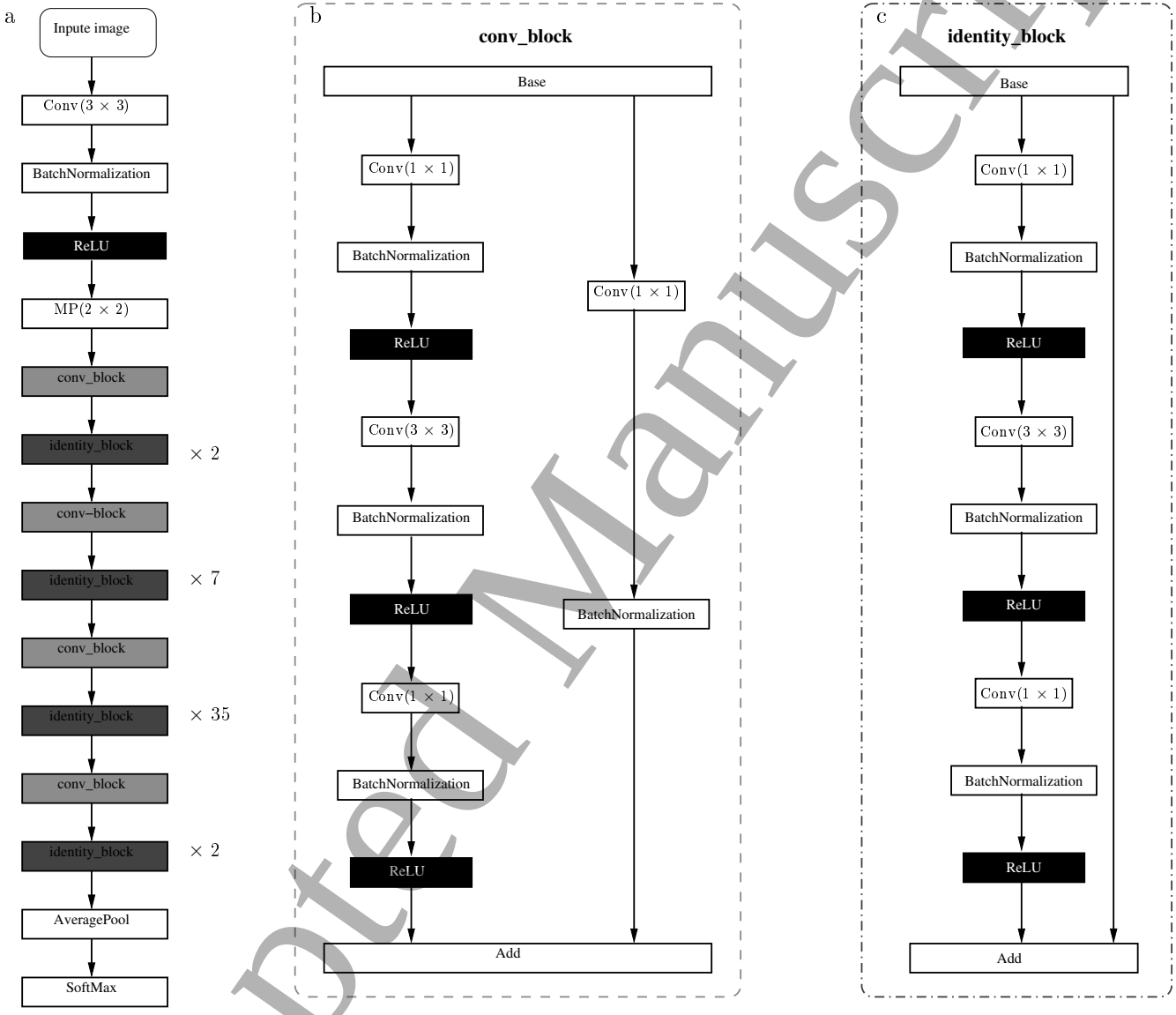


**Figure 7. Deep convolutional neural network architecture for classification of spatial patches based on biomimetic echoes**. **(a)** Overall architecture of the ResNet152 with four convolution blocks and 46 identity blocks), **(b)** architecture of an individual convolution block with three convolution stages and one layer convolution used to adjust the number of filters. **(c)** identity block architecture with three convolution layers and the original input propagated in parallel.

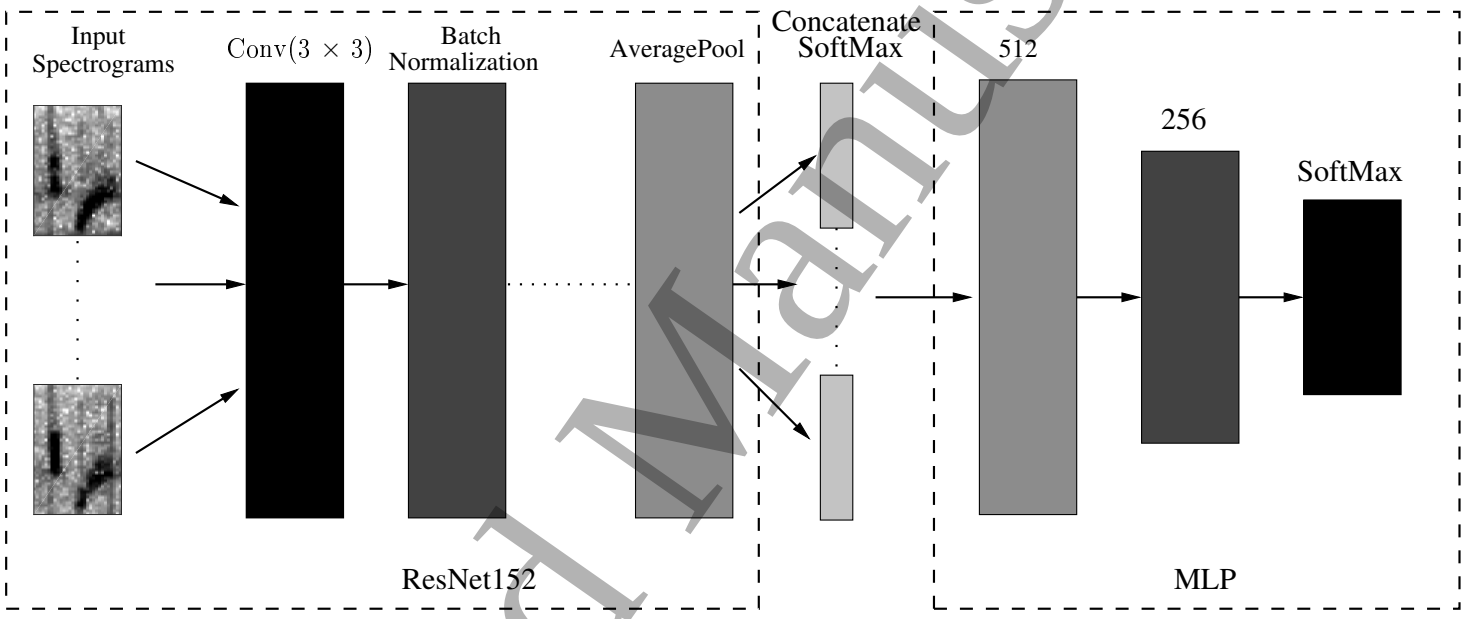


**Figure 8. Network architecture for the identification of spatial patches based on sets of multiple echoes**. The spectrogram representations of all echoes in the set are fed into a ResNet152 to extract time-frequency features from the entire echo set. The feature vectors derived from the output of the final SoftMax layer of the ResNet152 were concatenated into a single vector containing the feature maps for all individually echoes. The concatenated feature vector is passed into a multi-layer perceptron (MLP) to perform the supervised identification of the corresponding spatial patches.

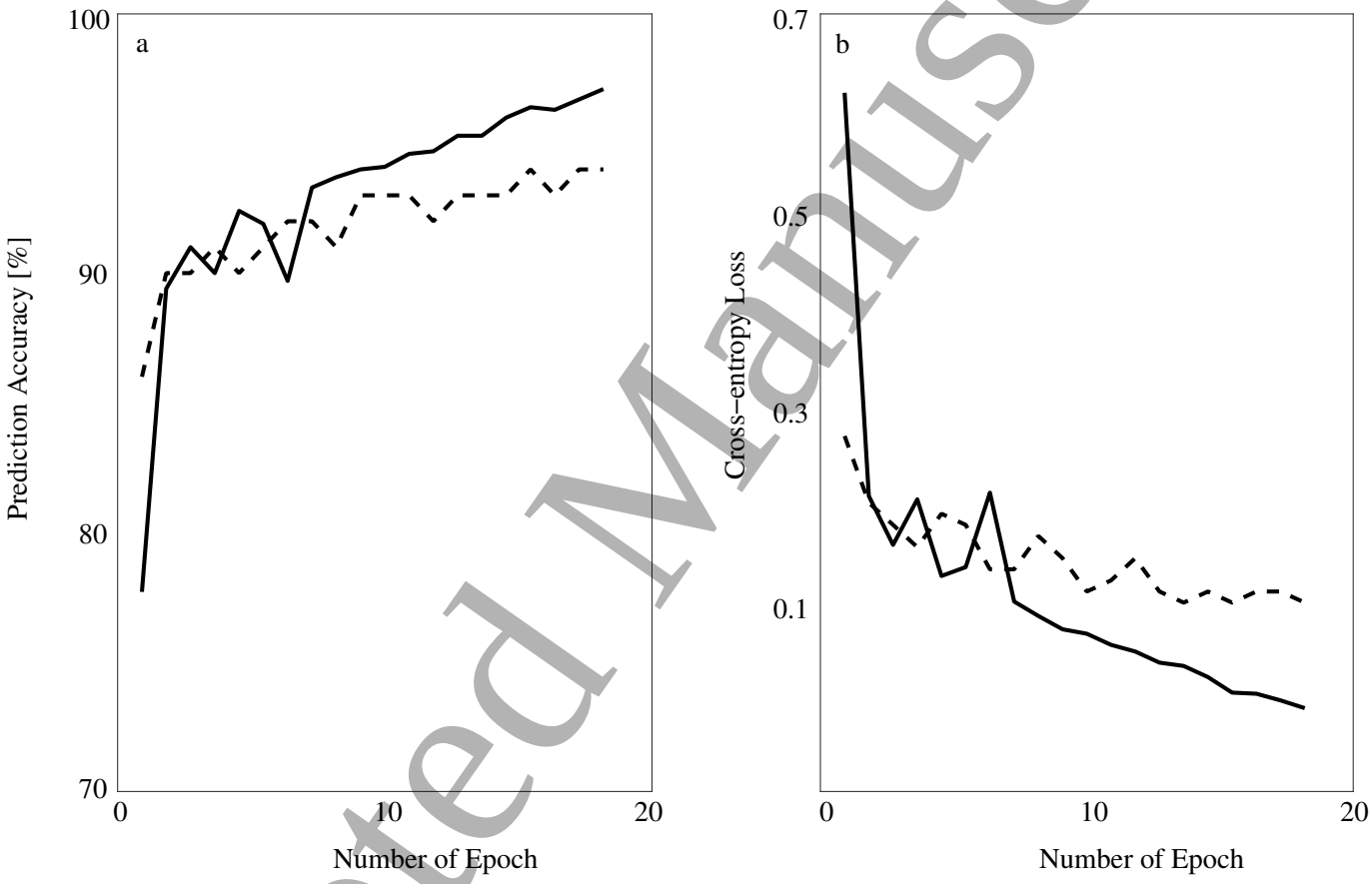


**Figure 9. Training (solid line) and validation (dashed line) performance of the deep neural network for location identification, one echo used to classify two patches**. **(a)** Prediction accuracy curve along the number of epochs. **(b)** Cross-entropy loss curve along the number of epochs.

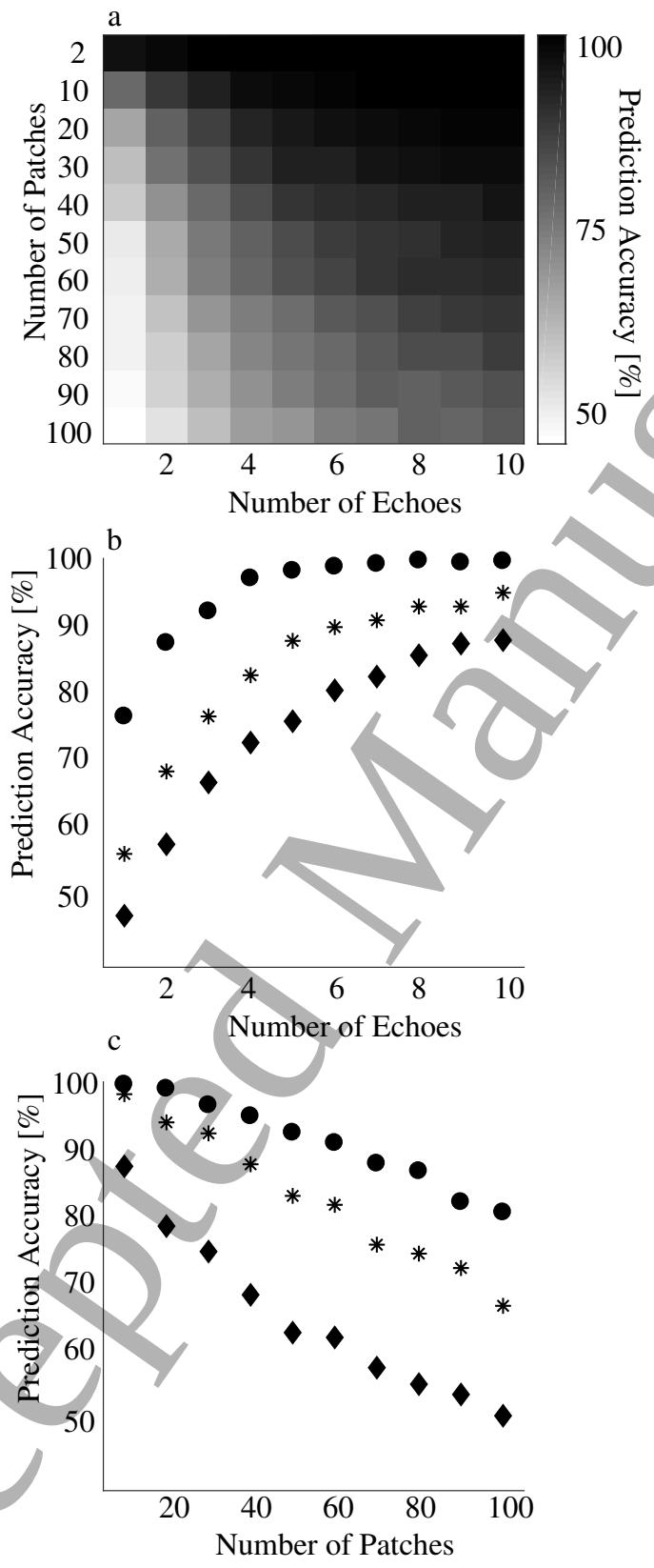


**Figure 10. Location identification performance for different numbers of spatial patches and echoes. (a)** Performance as a function of both variables (number of patches and echoes). **(b)** Prediction accuracy as a function echo data set size for three different number of spatial patches (circles: 10 patches, stars: 40 patches, diamond: 80 patches). **(c)** Prediction accuracy as a function of the number of spatial patches for different echo set sizes (diamonds: 2 echoes, stars: 5 echoes, circles: 10 echoes).

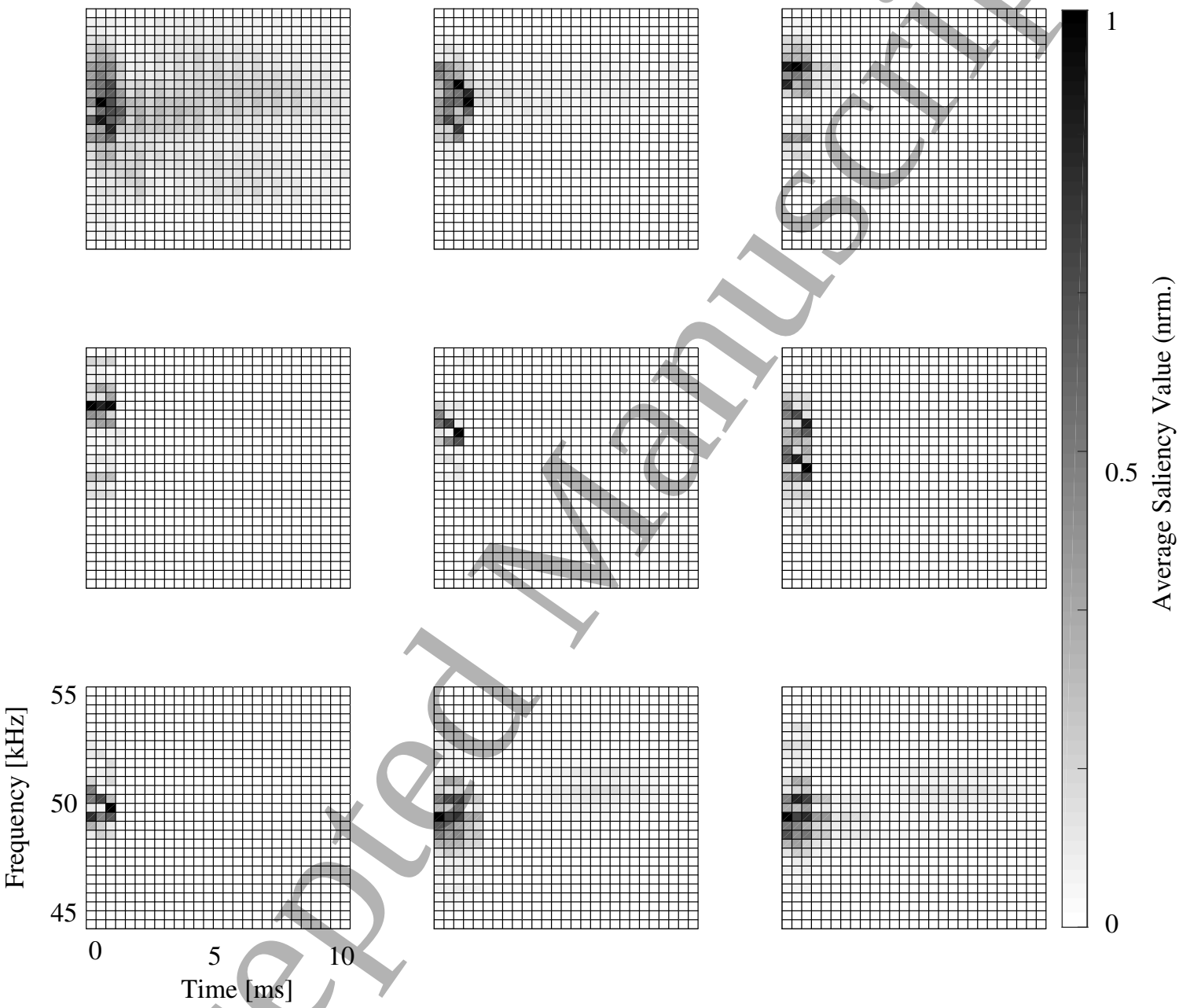


**Figure 11. Classification features in the time-frequency domain**. Average of 2,000 saliency maps for nine different spatial patches. The data set sizes for this figure ranged from 2,500 to 4,600 saliency maps. For data sets greater than 2,000, the averaged saliency maps were randomly picked to yield an equal sample size. Each saliency map has the same size as the input spectrogram.

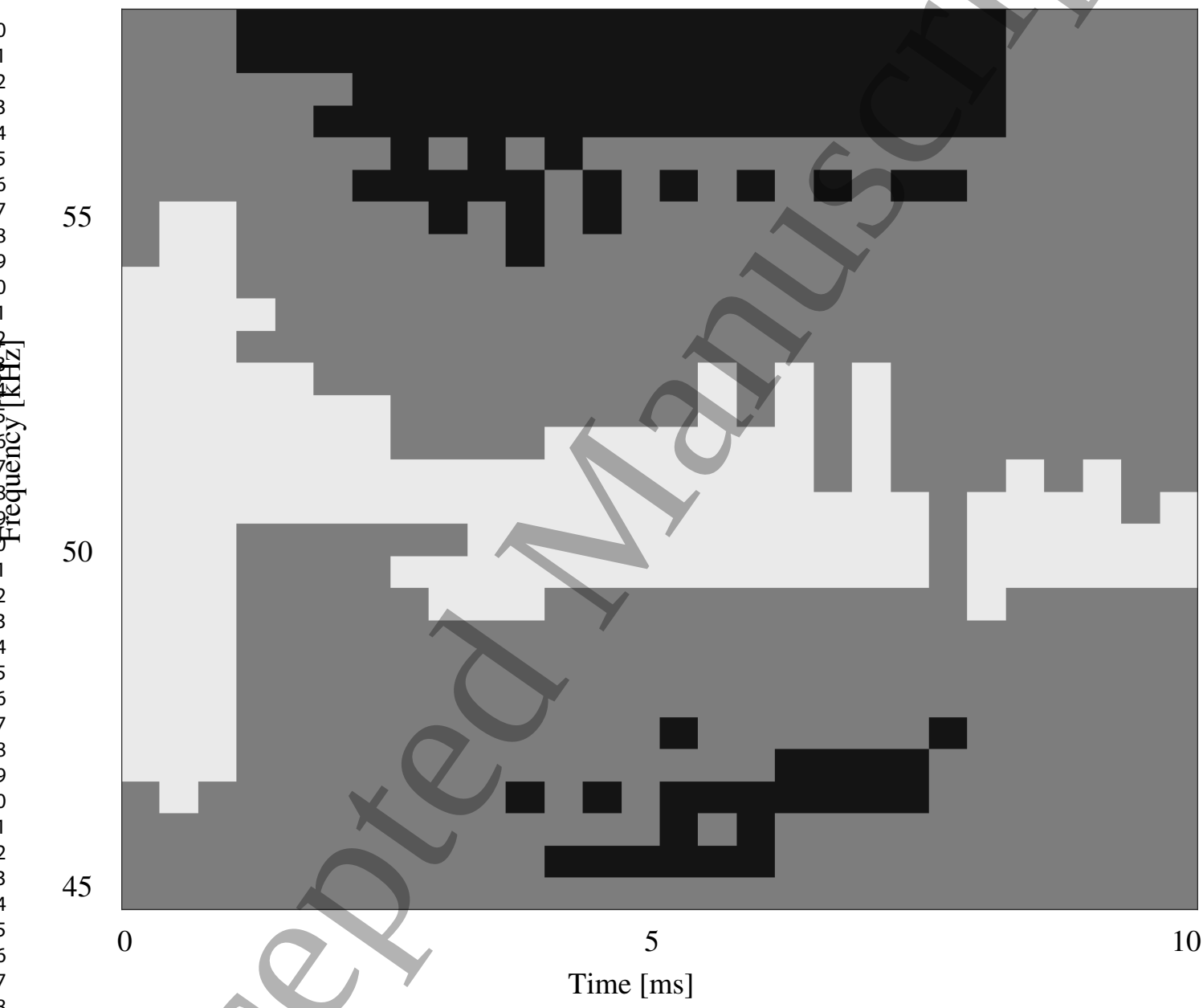


**Figure 12. Breakdown of the echo time-frequency plane into regions of different saliency**. Top 50% salience intersection (light gray), bottom 50% (black) . The regions were determined as the intersection of the individual saliency values, i.e., a time-frequency bin belongs to the top 50% values if the saliency values in all individual maps belong to that value range.

## References


1. Karma, S. *et al.* Use of unmanned vehicles in search and rescue operations in forest fires: Advantages and limitations observed in a field trial. *International journal disaster risk reduction* **13**, 307–312 (2015).
2. Burke, C. *et al.* Requirements and limitations of thermal drones for effective search and rescue in marine and coastal areas. *Drones* **3**, 78 (2019).
3. Lygouras, E. *et al.* Unsupervised human detection with an embedded vision system on a fully autonomous uav for search and rescue operations. *Sensors* **19**, 3542 (2019).
4. Hameed, I. A. Intelligent coverage path planning for agricultural robots and autonomous machines on three-dimensional terrain. *Journal Intelligent & Robotic Systems* **74**, 965–983 (2014).
5. Bac, C. W., Van Henten, E. J., Hemming, J. & Edan, Y. Harvesting robots for high-value crops: State-of-the-art review and challenges ahead. *Journal Field Robotics* **31**, 888–911 (2014).
6. Song, G., Yin, K., Zhou, Y. & Cheng, X. A surveillance robot with hopping capabilities for home security. *IEEE Transactions on Consumer Electronics* **55**, 2034–2039 (2009).
7. Hofmann-Wellenhof, B., Lichtenegger, H. & Collins, J. *Global positioning system: theory and practice* (Springer Science & Business Media, 2012).
8. Taraldsen, G., Reinen, T. A. & Berg, T. The underwater gps problem. In *OCEANS 2011 IEEE-Spain*, 1–8 (IEEE, 2011).
9. Leonard, J. J. & Bahr, A. Autonomous underwater vehicle navigation. *Springer handbook ocean engineering* 341–358 (2016).
10. Kalita, H., Morad, S., Ravindran, A. & Thangavelautham, J. Path planning and navigation inside off-world lava tubes and caves. In *2018 IEEE/ION Position, Location and Navigation Symposium (PLANS)*, 1311–1318 (2018).
11. Bakambu, J. N. & Polotski, V. Autonomous system for navigation and surveying in underground mines. *Journal Field Robotics* **24**, 829–847 (2007).
12. Grant, A., Williams, P., Ward, N. & Basker, S. Gps jamming and the impact on maritime navigation. *The Journal Navigation* **62**, 173–187 (2009).
13. Merry, K. & Bettinger, P. Smartphone gps accuracy study in an urban environment. *PloS one* **14**, e0219890 (2019).
14. Puricer, P. & Kovar, P. Technical limitations of gnss receivers in indoor positioning. In *2007 17th International Conference Radioelektronika*, 1–5 (IEEE, 2007).
15. Hsu, L.-T., Gu, Y. & Kamijo, S. Sensor integration of 3d map aided gnss and smartphone pdr in urban canyon with dense foliage. In *Proceedings of IEEE/ION PLANS 2016*, 85–90 (2016).
16. Prasser, D. & Wyeth, G. Probabilistic visual recognition of artificial landmarks for simultaneous localization and mapping. In *2003 IEEE International Conference on Robotics and Automation (Cat. No. 03CH37422)*, vol. 1, 1291–1296 (IEEE, 2003).
17. Kim, D., Lee, D., Myung, H. & Choi, H.-T. Artificial landmark-based underwater localization for auvs using weighted template matching. *Intelligent Service Robotics* **7**, 175–184 (2014).
18. Cope, J. S., Corney, D., Clark, J. Y., Remagnino, P. & Wilkin, P. Plant species identification using digital morphometrics: A review. *Expert Systems with Applications* **39**, 7562–7573 (2012).
19. Côté, J.-F., Widlowski, J.-L., Fournier, R. A. & Verstraete, M. M. The structural and radiative consistency of three-dimensional tree reconstructions from terrestrial lidar. *Remote Sensing Environment* **113**, 1067–1081 (2009).
20. Gézero, L. & Antunes, C. Automated three-dimensional linear elements extraction from mobile lidar point clouds in railway environments. *Infrastructures* **4**, 46 (2019).
21. Anand, B., Barsaiyan, V., Senapati, M. & Rajalakshmi, P. An experimental analysis of various multi-channel lidar systems. In *2020 IEEE International Conference on Computing, Power and Communication Technologies (GUCON)*, 644–649 (IEEE, 2020).
22. Heinrich, S. & Motors, L. Flash memory in the emerging age of autonomy. *Flash Memory Summit* 1–10 (2017).
23. Neuweiler, G. *et al.* Foraging behaviour and echolocation in the rufous horseshoe bat (rhinolophus rouxi) of sri lanka. *Behavioral ecology sociobiology* **20**, 53–67 (1987).
24. Genzel, D., Yovel, Y. & Yartsev, M. M. Neuroethology of bat navigation. *Current Biology* **28**, R997–R1004 (2018).

**25.** Meyer, C. F., Weinbeer, M. & Kalko, E. K. Home-range size and spacing patterns of macrophyllum macrophyllum (phyllostomidae) foraging over water. *Journal mammalogy* **86**, 587–598 (2005).

**26.** Reyer, H.-U. *et al.* Nectar intake and energy expenditure in a flower visiting bat. *Oecologia* **63**, 178–184 (1984).

**27.** Müller, R. & Kuc, R. Foliage echoes: a probe into the ecological acoustics of bat echolocation. *The Journal Acoustical Society America* **108**, 836–845 (2000).

**28.** Yovel, Y., Stilz, P., Franz, M. O., Boonman, A. & Schnitzler, H.-U. What a plant sounds like: the statistics of vegetation echoes as received by echolocating bats. *PLoS Computational Biology* **5**, e1000429 (2009).

**29.** McKerrow, P. & Harper, N. Plant acoustic density profile model of ctfm ultrasonic sensing. *IEEE Sensors Journal* **1**, 245–255 (2001).

**30.** Bhardwaj, A., Khyam, M. O. & Müller, R. Biomimetic detection of dynamic signatures in foliage echoes. *Bioinspiration & Biomimetics* **16**, 046026 (2021).

**31.** Zhang, L. & Müller, R. Large-scale recognition of natural landmarks with deep learning based on biomimetic sonar echoes. *Bioinspiration & Biomimetics* **17**, 026011 (2022).

**32.** Adafruit Industries. *Adafruit Ultimate GPS DataSheet* (2013). Rev. 3.

**33.** Dougherty, E. R. An introduction to morphological image processing. *SPIE, 1992* (1992).

**34.** Jones, G. & Rayner, J. Foraging behavior and echolocation of wild horseshoe bats rhinolophus ferrumequinum and r. hipposideros (chiroptera, rhinolophidae). *Behavioral Ecology Sociobiology* **25**, 183–191 (1989).

**35.** Newling, J. & Fleuret, F. Nested mini-batch k-means. *Advances neural information processing systems* **29**, 1352–1360 (2016).

**36.** Pedregosa, F. *et al.* Scikit-learn: Machine learning in python. *Journal machine learning research* **12**, 2825–2830 (2011).

**37.** Rousseeuw, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal computational applied mathematics* **20**, 53–65 (1987).

**38.** He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778 (2016).

**39.** Abadi, M. *et al.* TensorFlow: Large-scale machine learning on heterogeneous systems (2015). Software available from tensorflow.org.

**40.** Chollet, F. keras. https://github.com/fchollet/keras (2015).

**41.** Ioffe, S. & Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, 448–456 (PMLR, 2015).

**42.** Glorot, X., Bordes, A. & Bengio, Y. Deep sparse rectifier neural networks. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, 315–323 (JMLR Workshop and Conference Proceedings, 2011).

**43.** Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. & Wojna, Z. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2818–2826 (2016).

**44.** Goodfellow, I., Bengio, Y. & Courville, A. *Deep learning* (MIT press, 2016).

**45.** Jones, G. Scaling of wingbeat and echolocation pulse emission rates in bats: why are aerial insectivorous bats so small? *Functional Ecology* 450–457 (1994).

**46.** Schnitzler, H.-U. & Kalko, E. K. Echolocation by insect-eating bats: we define four distinct functional groups of bats and find differences in signal structure that correlate with the typical echolocation tasks faced by each group. *Bioscience* **51**, 557–569 (2001).

**47.** Fu, Z.-Y. *et al.* Sexual dimorphism in echolocation pulse parameters of the cf-fm bat, hipposideros pratti. *Zoological Studies* **54**, 1–9 (2015).

**48.** Riedmiller, M. & Lernen, A. Multi layer perceptron. *Machine Learning Lab Special Lecture, University Freiburg* 7–24 (2014).

**49.** Stone, M. Cross-validatory choice and assessment of statistical predictions. *Journal royal statistical society: Series B (Methodological)* **36**, 111–133 (1974).

**50.** Murphy, K. P. *Machine learning: a probabilistic perspective* (MIT press, 2012).

**51.** Adebayo, J. *et al.* Sanity checks for saliency maps. *arXiv preprint arXiv:1810.03292* (2018).

**52.** Müller, R. A computational theory for the classification of natural biosonar targets based on a spike code. *Network: Comput. Neural Syst.* **14**, 595–612, DOI: 10.1088/0954-898X/14/3/311 (2003).

**53.** Ucar, Z., Bettinger, P., Weaver, S., Merry, K. L. & Faw, K. Dynamic accuracy of recreation-grade gps receivers in oak-hickory forests. *Forestry: An International Journal Forest Research* **87**, 504–511 (2014).

**54.** Lee, T., Bettinger, P., Cieszewski, C. J. & Gutierrez Garzon, A. R. The applicability of recreation-grade gnss receiver (gps watch, suunto ambit peak 3) in a forested and an open area compared to a mapping-grade receiver (trimble juno t41). *PLoS One* **15**, e0231532 (2020).

**55.** Wang, R., Liu, Y. & Müller, R. Detection of passageways in natural foliage using biomimetic sonar. *Bioinspiration & Biomimetics* **17**, 056009 (2022).

**56.** Eisenberg, J. F. & Wilson, D. E. Relative brain size and feeding strategies in the chiroptera. *Evolution* 740–751 (1978).

**57.** Pritsker, A. A. B. *Introduction to Simulation and SLAM II* (Halsted Press, 1984).