Comparing Vision-based to Sonar-based 3D Reconstruction

Netanel Frank[1], Lior Wolf[1,2], Danny Olshansky[1], Arjan Boonman[1], and Yossi Yovel[1]
[1]Tel Aviv University, [2] Facebook AI Research

Abstract—Our understanding of sonar based sensing is very limited in comparison to light based imaging. In this work, we synthesize a ShapeNet variant in which echolocation replaces the role of vision. A new hypernetwork method is presented for 3D reconstruction from a single echolocation view. The success of the method demonstrates the ability to reconstruct a 3D shape from bat-like sonar, and not just obtain the relative position of the bat with respect to obstacles. In addition, it is shown that integrating information from multiple orientations around the same view point helps performance.
The sonar-based method we develop is analog to the state-of-the-art single image reconstruction method, which allows us to directly compare the two imaging modalities. Based on this analysis, we learn that while 3D can be reliably reconstructed form sonar, as far as the current technology shows, the accuracy is lower than the one obtained based on vision, that the performance in sonar and in vision are highly correlated, that both modalities favor shapes that are not round, and that while the current vision method is able to better reconstruct the 3D shape, its advantage with respect to estimating the normal's direction is much lower.

Index Terms—Sonar imaging, 3D reconstruction, hypernetworks

1 Introduction

Bats rely on both echolocation and on sight in order to sense nearby objects and to navigate in 3D space. Which of the two modalities is dominant, depends on the exact species. While most bats rely on echos, some fruit bats rely solely on sight.

As humans, we often assume that sight is preferable to echolocation, and that bats rely on the latter due to the darkness of their natural habitats or the need to hunt at night. As far as we can ascertain, no previous work has validated this assumption and directly compared the two modalities, which is the goal of this work.

In order to perform this comparison, we first construct a version of the ShapeNet benchmark that is based on sonar. In each 3D viewpoint, we simulate two adjacent bat-like ears. We then develop a method for reconstructing the 3D shape given the sonar signal. Several architectures are compared and we rely on a hypernetwork scheme, in which the perceptual analysis is performed by a convolutional recurrent neural network, and the shape itself is represented implicitly by a classification network.

We then use the developed tools in order to directly compare with a ShapeNet reconstruction method that is based on a single 2D image. The reason we compare with a single view despite using two ears, is that we are more interested in understanding the bat capabilities, and one ear is not analog to one eye. To understand the last point, consider that a single ear provides distance, the second ear adds a single angle (azimuth) based on the time difference, i.e. two degrees of freedom, just as the case with a single image that provides two angles (azimuth and elevation). In addition, previous work on Shapenet has shown that multiple viewpoints, even if well separated (unlike the small baseline of a typical stereo pair) only contribute marginally on this dataset [1].

The comparison between the modalities is done to a similar implicit shape method, in which the shape is represented by a classifier that for every point in 3D determines whether it is inside or outside the shape. Specifically, the image based model employs a ResNet-based hypernetwork, i.e., the weights of the network that classifies 3D points are provided by the ResNet that processes the input image.

The similarity of the methods used in the two modalities allows us to perform a rather direct comparison and come up with the following conclusions:

- Vision leads, in most cases, to more accurate reconstruction, at least under the studied conditions.
- Typically, examples that are easy to reconstruct based on one modality are also easy to reconstruct by the other modality.
- In shapes where both sonar and vision methods work well, the advantage of vision over sonar is much greater in terms of occupancy in voxel space than in terms of estimating the normal's direction.
- Sonar may have an advantage over vision around the shape's corners. However, both modalities prefer shapes that are not round.

2 Related Work

Fascinated by the bats' ability to sense the world acoustically, many attempts have been made to model and mimic their abilities. Using either real-recorded or simulated echoes, several previous studies aimed to classify echoes and characterize their statistics [2]. A wide range of methods were applied either to the echo's time, spectral or tempo-spectral domains using different approaches including: extraction of sets of acoustic features and using statistical tests [3], [4], modeling echo formation [5], [6] and biological processing [7] and using various types of machine learning algorithms (e.g. SVMs as in [8]). A few recent attempts

also used artificial neural networks for echo-based object recognition and estimation [9].

Bats' ability to reconstruct 3D using echoes remains a fundamental open question. There is currently a debate on whether bats reconstruct 3D to perceive the environment, or simply use echoes statistically, that is, without localizing the multiple reflectors generating them (or combine both). There is also a debate how many echoes would be required for 3D reconstruction. In this study, we address these questions empirically, by training a neural network to reconstruct 3D from a single echo.

2.1 3D reconstruction from single views

The availability of large scale CAD-based datasets, such as ShapeNet [10] have led to a proliferation of deep-learning based methods for 3D reconstruction. The deep learning architecture is largely affected by the 3D representations that are utilized and the literature can be divided into four categories: (i) voxel based methods, which are 3D grids that generalize pixels, (ii) polygon meshes and similar topology preserving representaitons, (iii) point clouds and similar representations of densities in 3D, and (iv) implicit surfaces.

Since the technology to generate images as pixel grids is highly evolved, voxel based methods are highly popular. However, due to their cubic memory to resolution ratio, voxel solutions suffer from limited resolutions. To overcome this, nested architectures, such as Octrees, have been used [11], [12], [13], [14].

Mesh based representations were used in concert with a differential renderer within an analysis by synthesis framework [15], [16]. In another work, a graph neural network was used to generate a mesh [17].

Point clouds form an efficient and scaleable representation, with the disadvantage of having to reconstruct the topology before rendering the 3D model. From the technical point of view, generating a set of unordered entities of an arbitrary size may challenge training. This is overcome by adding stochasticity to the solution and designing appropriate loss functions [18] and by relying on the points of the input image as a guide [19].

The implicit field shape representation employs a classifier to define the shape. This classifier is conditioned on an embedding of the input image given by a learned encoder [20], [21], [22]. The "decoder" receives a 3D coordinate as well as the embedding vector, and the integration of both inputs require the usage of a relatively large network. The methods, therefore, suffer from very long train times. It is also not clear how these methods generalize to very large training sets which include multiple shape classes, since none of these publications have reported results on the commonly used ShapeNet ground-truth annotations and instead opted with retraining the baseline methods on subsets of the data. Mescheder et al. [22] is the only implicit-shape method to report (limited) multi-class results on a dataset that is derived from ShapeNet, but introduced additional supervision in the form of pre-training on imagenet.

2.2 Meta functionals [1]

The term Hypernetwork is now commonly used to refer to a technique in which one network f predicts the weights of another network g, which is called the primary network. During training, the weights of g are not learned directly, as these are generated by f depeding on the latter's input.

The first contributions to employ such a scheme relied on a specific dynamic convolutional layer in order to transform the input in an adaptive manner [23], [24]. Networks with more dynamic layers were subsequently used for video frame prediction [25]. The term hypernetwork was coined in a work that studied RNNs in the context of NLP [26]. Since f can transfer information between related tasks, g can adapt between tasks and hypernetworks, therefore, are especially suitable for few-shot learning [27].

Our sonar reconstruction method employs the general high level scheme recently proposed by Littwin and Wolf for 3D reconstruction from a single image [1]. This scheme, termed meta functionals, can be seen as a hypernetwork version of the implicit shape representation [20], [21], [22]. However, while the latter methods struggle to produce competitive results on ShapeNet, meta functionals are currently the state of the art method in 3D reconstruction from a single image.

The meta functional hypernetwork has two subnetworks f, g with parameter values θ_f, θ_I respectively. The weights θ_f are learned during the training phase. The weights of the primary network g are a function of I, the input image at hand, and are produced by the network f.

Similar to the implicit shape work, g is a classification function that maps a point p with coordinates (x, y, z) in 3D into a score $s_I^p \in [0, 1]$, such that the shape is defined by the classifier's decision boundary, i.e., it is a mapping from a vector in $\mathbb{R}^3$ to a scalar. f maps between completely different domains: the input image I to the parameters θ_I of network g, and utilizes a deep ResNet.

This hypernetwork scheme is given by:

$$\theta_I = f(I, \theta_f) \tag{1}$$

$$s_I^p = g(p, \theta_I) \tag{2}$$

In the single image reconstruction task, $f(I, \theta_f)$ is a deep ResNet. The primary network $g(p, \theta_I)$ is a Multi-Layered Perceptron (MLP) with four layers, in most experiments.

Since hypernetworks often struggle at initialization, the parameterization of the weights of g slightly differs from the conventional parameterization. Each layer n of the MLP g performs the following computation:

$$y = ((\theta_I^{W(n)} x) \cdot \theta_I^{s(n)}) + \theta_I^{b(n)} \tag{3}$$

where $\cdot$ denotes the Hadamard product, x is the layer's input, y is the layer's computation result, $\theta_I^{W(n)}$ is the weight matrix of layer n, $\theta_I^{b(n)}$ is the bias vector of that layer, and $\theta_I^{s(n)}$ is the learned scale vector. θ_I, which is the output of f, is a concatenation of $\theta_I^{W(n)}$,$\theta_I^{b(n)}$ and $\theta_I^{s(n)}$ for all layers $n \in [1, 4]$.

Note that the two networks f, g define a directed acyclic graph and, therefore, the backpropagation algorithm can be naturally applied (and is also easy to apply with modern deep learning frameworks, which employ automatic differentiation).

The loss is the binary cross entropy loss between $s_I^p \in \mathbb{R}$ (Eq. 2) and $y(p) \in \{0, 1\}$, which is the ground truth label of whether the point p is inside ($y(p) = 1$) or outside ($y(p) = 0$)

the 3D shape. Formally, given the image I and the ground truth shape y, the the loss is a function of the parameters of f

$$H(\theta_f, I) = -\int_V y(p)\log(g(p, f(I, \theta_f))) + (1 - y(p))\log(1 - g(p, f(I, \theta_f)))dp \quad (4)$$

where V is the 3D volume in which the shapes reside. This is minimized over θ_f across all pairs of images and 3D models in the training set and over points that are sampled from the volume V.

3 METHOD

We present a method for shape representation using bat-like echoes. First, we create EchoNet — a dataset of synthesized echoes coupled with the corresponding 3D meshes that were used to create the signals. We then develop BatNet — a deep meta functional architecture for the reconstruction of 3D objects from a pair of echoes.

3.1 Data Generation

EchoNet was synthesized using an acoustic simulator that calculates the acoustic impulse response of a 3D mesh. The 3D objects were taken from ShapeNet [10]. Each 3D mesh file was processed using the acoustic simulator in order to create two echoes at a small distance apart (baseline of 2 cm) around a specific point of view. Similar to the image rendering of ShapeNet, the echoes were produced at 24 different points of view, equally spaced on a circle at an elevation angle of 30 degrees, at a radius of 3m from the model center (the model size is normalized so its longest axis will be 1 meter long). As can be seen in Fig. 1, the sonar sensors (ears) can be oriented arbitrarily around each point of view (location of the emitter). This is unlike the computer vision dataset, where the views, in most ShapeNet benchmarks, are taken such that at least one axis is aligned.

As a requirement for the echo generation pipeline, the input to the acoustic simulator should be a watertight 2-manifold mesh. The 3D objects that were used do not necessarily meet this criterion. In order to comply, each mesh file was preprocessed using the method described in [28], which generates a watertight triangular mesh. The model was then simplified, thus keeping the acoustic simulator run-time to the necessary minimum.

Acoustic Simulator The acoustic simulator is based on an approximation to the Boundary Element Method [29]. Instead of solving the computationally expensive boundary equations for each boundary element (mesh face), a raytracing like computation is performed [30]. The simulator was compared to real object echoes and was found to be highly accurate (see Supplementary figures in [30]).

For each frequency in the required range, a sum is calculated over the responses of all the faces. For every face, the response is the complex superposition of the two-way free space propagation loss, with the reflection loss from the face. The reflection loss changes with the angle from the point of view to the face's normal. The reflection loss of a face (as a function of the impact angle and frequency) had been calculated in advance using a full boundary element model (BEMFA). The simulator only selects faces whose normals point toward the hemisphere containing the point of view. It also eliminates reflecting faces that are occluded by other parts of the object.

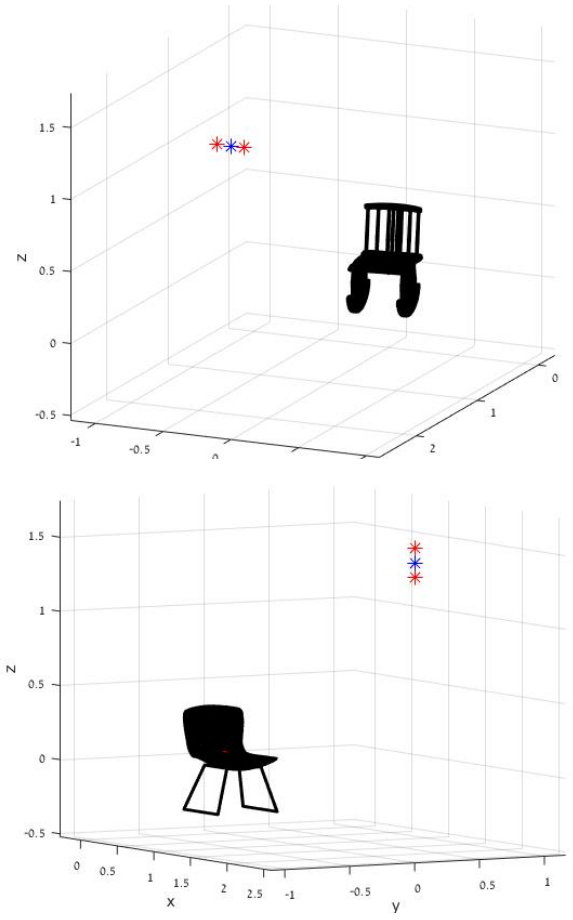

Fig. 1. The location of the sonar sensors in 3D with respect to the object, In blue - the location of the emitter, in red - the two "ears". Unlike the common vision-based Shapenet views, the axes of the sensor are not aligned with the world coordinate system in any way.

Using the inverse Fourier transform, the object's acoustic impulse response is then acquired. Fig. 2 depicts an example of the output of the steps of the process of capturing a 3D object.

3.2 The sonar hypernetwork

The architecture of BatNet is based on deep meta functionals shape representation [1]. The method employs two networks f and g. f can be seen as an encoder that maps an input echo E to an embedding θ_E, which is directly utilized as the weights of network g for the reconstruction of the shape that is captured in E. g is an MLP network that classifies a 3D point p with coordinates (x, y, z) into a score $s_E^p \in [0, 1]$. Similar to Eq. 1, 2 this relation can be described by:

$$\theta_E = f(E, \theta_f) \quad (5)$$

$$s_E^p = g(p, \theta_E). \quad (6)$$

$s_E^p \in [0, 1]$ is the score that determines whether a point p is likely to be inside (values closer to one) or outside the

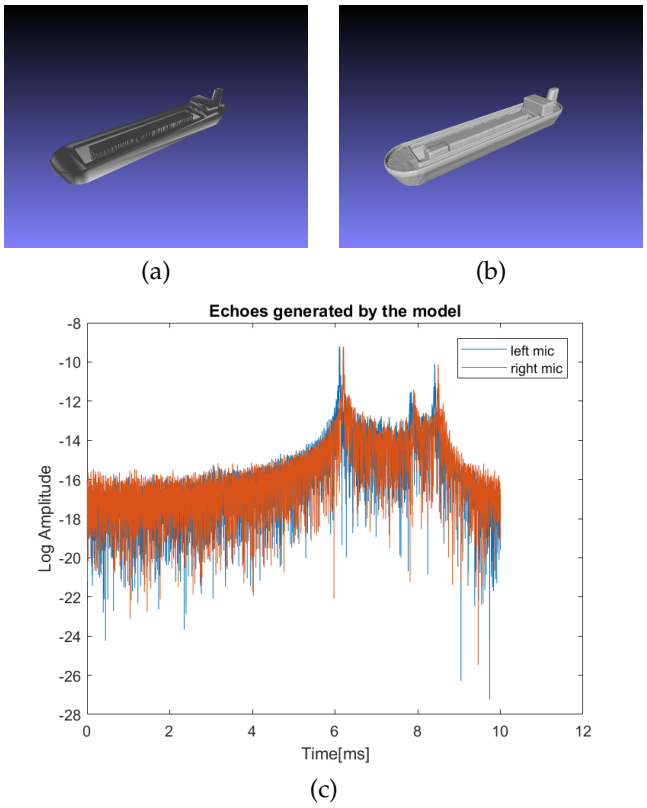

Fig. 2. The input and the output of the acoustic simulator. (a) the original ShapeNet model, (b) the processed watertight model, and (c) the two channel output of the acoustic simulator.

shape (values closer to zero). The shape's surface is defined by the decision boundary of s_E^p .

In [1], it was shown that the MLP implementation for g is capable of representing the 3D object well. It was also shown that the exact architecture of g has little effect. Our implementation employs four layers, with 32 hidden neurons each, and ELU activations [31]. These are the default parameters used in [1], which allows us to directly compare with the shape representation employed in their image-based hypernetwork.

Where vision and echolocation differ is, of course, in the input signal. We, therefore, focus on the different architectures of the encoders employed for the weight generating network f, looking for the best approach to capture the information in the echos. Two main approaches for input representation were evaluated, the first using the 1D 2-channel acoustic signal (wav) and the second using the Short Time Fourier Transform (STFT) of the signal.

3.3 Architectures

As for the different input representation methods, three encoders (network f) were considered.

3.3.1 SoundNet

For the case of using the acoustic signal as the network input, SoundNet [32] was chosen as the encoder; Changes were made to the hyperparameters in order to accommodate the smaller signal size. The encoder was applied to the two channels separately, and then the two embeddings were concatenated to create a single embedding vector. A single linear layer was added to regress to θ_E.

The SoundNet model was constructed similarly to the original 5-layer architecture. It consists of four 1D-conv layers with stride 2, the base layer has 32 filters of size 64, each consecutive layer increases the number of filters by two and decreases the filter size by the same factor. The first three layers are followed by a max pool of size 2, instead of 8 in the original implementation, the last layer is a 1D-convolution layer with 1401 filters of size 16 and stride 2 (instead of 12). After the two channels are concatenated, it passes through two fully connected layers of size 1024 each.

3.3.2 ResNet

For the STFT version, a few encoders were tested. The base architecture employs two spectrograms as inputs, or, more specifically, the absolute value of the Echo's STFT of the two sensors. This two-channel input is processed by five ResNet blocks, followed by two fully connected layers.

An alternative architecture is based on the same network architecture but it also utilizes the phase of the STFT. Thus, the input to the second variant is a 4-channel image with two spectrograms and their phases.

In both cases, the Resnet encoder is based on the ResNet-34 model [33]. The input conv layer has 64 filters with a 5×5 kernel and stride 1, followed by five residual blocks with three layers each employing a 3×3 kernel. Every block increases the number of filters and reduces the spatial resolution by a factor of 2. Each convolutional layer is followed by batch normalization and ReLU. The output of the ResNet blocks is passed to a double fully connected layer of size 1024.

3.3.3 BatNet

Finally, the chosen architecture, we term BatNet, employs the same four channel input, but with a convolutional recurrent neural network (CRNN) [34] encoder, as previously adapted to audio classification by SELDNet [35], which was designed for multi-channel sound data represented as spectrogram and phase images.

The encoder of BatNet consists of blocks of convolutional layers, followed by bi-directional GRUs, the GRUs are followed by fully connected layers, as can be seen in Fig. 3. The BatNet architecture utilizes the time-frequency duality of the spectrogram by preserving the time resolution, while reducing the frequency resolution using an unequal Max-Pooling strategy, then passing the resulted time series of the extracted features through a bi-directional GRU.

Similar to what was suggested by [35], BatNet architecture consists of five convolution layers, each with 64 filters of size 3×3 and stride 1. Each convolution layer is followed by a ReLU activation, batch normalization and uneven max-pool of size 1×2 . After the convolution block, the features are reshaped, according to the unchanged axis, to a time series that is fed to a double bi-directional GRU of size 128. The output of the GRU block is passed to a double fully connected layer of size 1024. Our implementation of the encoder differs from SELDnet in a few details: (i) we use five convolution blocks instead of three, (ii) our max pooling size is constant in all of the blocks, and (iii) we have two fully

Fig. 3. The full architecture of BatNet. The weight generating network f consists of a CRNN encoder. The primary network g is an MLP, which given a point p_{xyz} in 3D returns the prediction of whether it is inside or outside the shape.

connected layers of size 1024, instead of one such layer with 128 hidden units.

4 Experimental Results

Since there is no literature baseline to compare our results to, we show the results of several alternative echo based networks. To allow an analysis and comparison of the echo-based results to vision-based ones, we use the data split Choy et al [36] of ShapeNet. We further split the train set to a train-validation set using a ratio of $80\% - 20\%$.

4.1 Training parameters

During training, a set of 10K pre-sampled points were used as an input for g . In order to sample more points around the boundary surface, which will be more informative in training, points were taken from multiple size models, the original scale, and two other scales, one larger one smaller by approximately 3% of the original, see Fig. 4. During training, a white Gaussian noise was added to the points, and, in addition to the boundary points, another 1k uniformly distributed points in 3D were also included.

The SoundNet and ResNet-based models were trained for approximately $180K$ iterations using Adam optimizer with a batch size of 64 and a learning rate of 10^{-5} , the BatNet architecture was trained similarly but started with a learning rate of 5×10^{-5} which was dropped during the last two epochs to 10^{-5}.

Fig. 4. Point sampling in three scales, as can be seen on an airplane model.

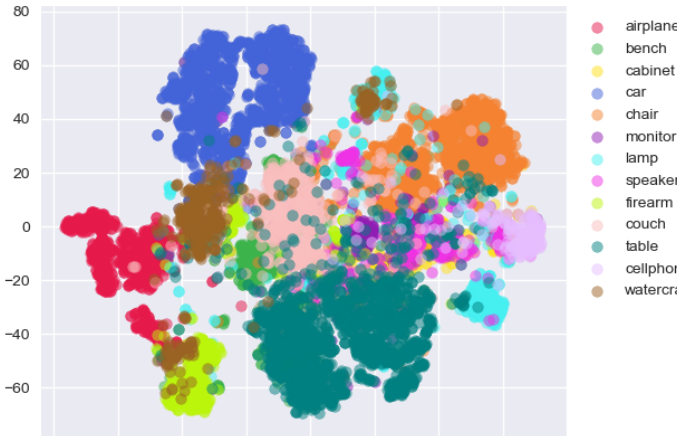

Fig. 5. Visualization of the learned embedding of the 13 different categories using t-SNE.

4.2 Quantitative results

The results of the different architectures are presented in Tab. 1. The first two results in the table are of leading image based architecture. The rest are based on echos.

The SoundNet based model achieves a mean IOU of 43.1%. This is 1.3% more than the base ResNet model, which uses only the absolute value of the STFT without the phase, and achieves the lowest performance of all the implementations that were tested - a mean IOU of 41.8%. The fact that the exact same architecture, but with the phase data added to the network input, achieves a mean IOU of 49.6%, suggests that the phase holds additional data that has a major contribution to the ability of the network to reconstruct the 3D model.

BatNet, with the time-freq influenced architecture, was trained and tested on both horizontal and vertical data (as seen in the lower image of Fig. 1), both achieved highest results of 52.0% and 51.6% accordingly, which shows the strengths of the architecture. In addition to obtaining the highest mean IOU results for echos, BatNet models also achieved the best per class IOU in 12/13 classes.

TABLE 1
Reconstruction results (IOU) for the different networks. Bold indicates the best out of the sonar networks.

Networks	Airplane	Bench	Cabinet	Car	Cellphone	Chair	Couch	Firearm	Lamp	Monitor	Speaker	Table	Watercraft	Mean
Images - 3D-R2N2 [36]	51.3	42.1	71.6	79.8	66.1	46.6	62.8	54.4	38.1	46.8	66.2	51.3	51.3	56.0
Images - Meta functional [1]	71.4	65.9	79.3	87.1	79.1	60.7	74.8	68.0	48.6	61.7	73.8	62.8	65.4	69.1
Echo -SoundNet	48.9	24.3	52.6	75.1	60.2	29.9	50.0	46.7	21.5	37.3	46.4	29.8	37.9	43.1
Echo – base ResNet	51.5	27.8	49.4	72.3	59.9	28.3	46.2	46.3	19.4	33.7	42.8	27.8	37.5	41.8
Echo – ResNet+phase	51.7	38.9	58.5	**78.9**	63.2	37.8	55.9	53.9	29.6	37.9	53.0	38.9	46.1	49.6
Echo – SELDnet original	50.2	32.3	48.8	71.3	63.1	30.7	51.3	55.1	25.5	40.4	43.2	32.3	43.7	45.2
Echo - BatNet - Vertical	55.2	**42.0**	59.6	78.7	67.4	**39.1**	58.2	54.6	**30.1**	42.0	**53.5**	**41.3**	**48.7**	51.6
Echo - BatNet - Horizontal	**56.4**	41.1	**60.4**	74.5	**71.7**	38.4	**58.9**	**56.1**	29.7	**52.2**	52.3	36.1	47.7	**52.0**

As mentioned, the BatNet architecture is based on that of SELDnet. It is evident from Tab. 1 that the original SELDnet obtains a considerably lower IOU (45.2%).

In Fig. 6 the per class IOU - distribution of vision ("Images - Meta functional") vs BatNet is presented. One can observe that a correlation exists between the IOU results of the two modalities. The Pearson correlation is R=0.97 when correlating the mean IoU of the image-based and the audition-based hypernetworks. When correlating at the single test sample level, the correlation is also very high (R=0.75). These numbers are remarkable given that the modalities are vastly different and that the data went through some preprocessing prior to the sonar signal being computed.

Although from Tab. 1 it is clear that the overall per class IOU is always better using vision, from the scatter results that are summarized in Tab. 2, we can see that BatNet is able to achieve as good or better results than the vision hypernetwork on 25.5% of the models. Examples for BatNet successful reconstruction over vision are presented in Fig. 8.

4.2.1 Normal direction distribution

In order to check the capability of the different modalities, vision-based deep functionals and BatNet, to reconstruct the normal's direction accurately, we focus on a subset of the models with a high IOU (over 0.7) in both modalities. For those models, we computed the shape's normal for each face, and consider the two angles that represent the normal direction Φ and Θ. Since matching the 3D model to the reconstructed one is a challenging problem by itself, we constructed 2D histograms to capture the distribution of the normal's direction.

Once obtained, it is possible to compare, using a correlation score, the resemblance of the histograms produced by the network of each of the modalities to the histogram of normals that was created using the original model. We define this correlation value as Normal-Corr. By comparing the number of times BatNet wins (achieve better results) based on the Normal-Corr to the number of wins using IOU values for the same set of models, we can verify which network is able to generate the normals more intuitively.

As shown in Fig. 7 BatNet achieves far more wins (or draws) with Normal-Corr than with IOU. From the summary in Tab. 3 we can observe that for 11/13 classes the win ratio for BatNet using Normal-Corr vs IOU is more than double. For four classes, BatNet has more wins using Normal-Corr than the vision based network.

4.3 Qualitative results

First we would like to verify that the BatNet encoder is capable of distinguishing between the different classes in an unsupervised manner, despite training a single network f for all classes, without conditioning on the class in any way. To do that, we employ t-SNE [37] in order to capture in 2D the activations of the penultimate layer of f (1024 dimensions) of the test data.

The results are depicted in Fig. 5. It is evident that the encoder is able to separate most of the 13 classes well. Second, the most difficult to learn class, "lamp", as can be seen from the IOU value in Tab. 1, is spread over a large part of the graph and divided into multiple sub groups. This implies that the class is not uniform, and thus hard to generalize and reconstruct.

4.4 Jacobian norm

We wish to evaluate the uncertainty of the network's boundary reconstruction. This can be done by examining the value of the network's gradients norm calculated at the 3D shape boundary, which is the norm of the Jacobian of the network w.r.t the 3D point p_{xyz}.

$$J_{xyz} \;=\; \frac{(\partial(g(p_{xyz},\ f(E)))}{(\partial\ p_{xyz}\ )} \tag{7}$$

Examples of different jacobians extracted using BatNet and vision networks is presented in Fig. 9. A yellower color represents a higher Jacibian norm, which is equivalent to a sharper decision boundary and lower uncertainty.

Qualitatively, (see examples in Fig. 9) it seems that the BatNet displays a higher certainty for 3D corners in comparison to the vision network. This is probably due to the fact that BatNet perceives normals more direclty than the overall shape, and corners are the singularity points of the normals.

However, verifying this quantitatively did not show different patters between vision and audition and both the visual and acoustic networks showed increased errors in classes of objects that are more round (have fewer corners). For this purpose, the roundness of the object was estimated

TABLE 2
Model based BatNet Vs. Vision IOU.

Networks	Airplane	Bench	Cabinet	Car	Cellphone	Chair	Couch	Firearm	Lamp	Monitor	Speaker	Table	Watercraft	Mean
Vision Wins [%]	81.5	74.6	78.8	81.6	62	82.1	76.4	74.6	68.0	66.8	76.8	73	72.8	74.5
BatNet Wins [%]	8.5	12.1	10.1	7.8	17	8.4	10.7	11.4	14.8	16.4	10.9	13.0	13.0	11.9
Even [%]	10	13.3	11	10.6	21	9.5	12.9	14	17.2	16.8	12.3	14	14.2	13.6

TABLE 3
IOU Vs. Normal-Corr BatNet wins ratio for models with IOU > 0.7.

Networks	Airplane	Bench	Cabinet	Car	Cellphone	Chair	Couch	Firearm	Lamp	Monitor	Speaker	Table	Watercraft	Mean
BatNet Wins IOU [%]	6.49	19.05	14.04	8.59	19.15	14.55	15.38	13.95	13.79	23.35	9.3	18.25	25	15.45
BatNet Wins Normal-Corr [%]	48.92	14.29	31.58	23.38	57.45	30.91	33.14	34.88	34.48	64.71	25.58	23.81	57.14	36.94
Normal-Corr to IOU win ratio	7.54	0.75	2.25	2.72	3.00	2.12	2.15	2.50	2.50	2.77	2.75	1.30	2.29	2.67

TABLE 4
BatNet IOU results, using one or three pairs of ears for the reconstruction.

Networks	Airplane	Bench	Cabinet	Car	Cellphone	Chair	Couch	Firearm	Lamp	Monitor	Speaker	Table	Watercraft	Mean
Echo - BatNet	56.4	41.1	60.4	74.5	71.7	38.4	58.9	56.1	29.7	52.2	52.3	36.1	47.7	52.0
BatNet - 3 pairs	61.9	49.7	67.4	78.2	76.9	46.3	65.2	62.4	37.2	55.8	61.8	48.3	57.3	59.1

by considering the variance of the normal directions in the ground truth data.

As can be seen in Fig. 10, there was a significant negative correlation between the average roundness of the object (which was assessed using the variance of point-norms) and the network's performance (IOU) in six classes ('lamp', 'speaker', 'couch', 'table', 'cellphone', 'watercraft'); Pearson correlation, $p < 0.05$ after a Bonferroni correction. Globally, over all test samples, the Pearson correlation is -0.12 for audition and -0.05 for vision. In the case of sonar, this negative correlation is probably due to the difficulty of measuring the distance to a round surface. In the case of vision, the cause may be that corners (edges) provide more spatial information than round surfaces.

4.5 Multiple sensor pairs

The BatNet experiments above utilize two echoes that were created using a single pair of "ears", as shown in Fig. 1. It is, however, possible to record the echos from additional angles for a single point of view, i.e. setting the point of view (the emitter location) and measuring the echos using multiple pairs of ears at different orientations relative to the object. This, to some approximation, mimics the integration of information that a moving bat can obtain, since the translation of the emitter is slower than the head's rotation.

Technically, the multi sensor pairs network is identical to the single pair BatNet, except that the number of input channels has increased: in the new network, the input consists of the data of the three pairs concatenated along the channel dimension.

As shown in Table. 4, using BatNet with 3 pairs of ears, it is possible to improve the mean IOU by 7.1%. The improvement is due to the fact that the rotation of the ears produces sampling along a different plane, which adds data to the reconstruction.

5 Conclusion

Computational models have been used in order to shed light on biological vision systems and a large body of literature has evolved specifically on the topic of comparing deep neural networks to vision in primates. However, the potential of deep learning in promoting the understanding of other sensory modalities and species that are more distant from us has been largely untapped.

Bat echolocation is a prime example of such a sensory system that could benefit from the revolution in deep learning. Many questions regarding the processing of echolocation echoes remain open, such as the fundamental question regarding bats' ability to build a 3D image of the world from the acquired echoes. Using deep learning on input data that is restricted to what is available to the bat, can shed light on what type of information can be extracted from this input.

In order to address these fundamental questions in echolocation processing , we design a sonar-based 3D reconstruction, such that it is directly comparable to the state-of-the-art single image 3D reconstruction networks. By comparing the performance of the two models, we come to multiple conclusions. First, echoes can be used to reconstruct the 3D shape of an object with surprising performance that, in

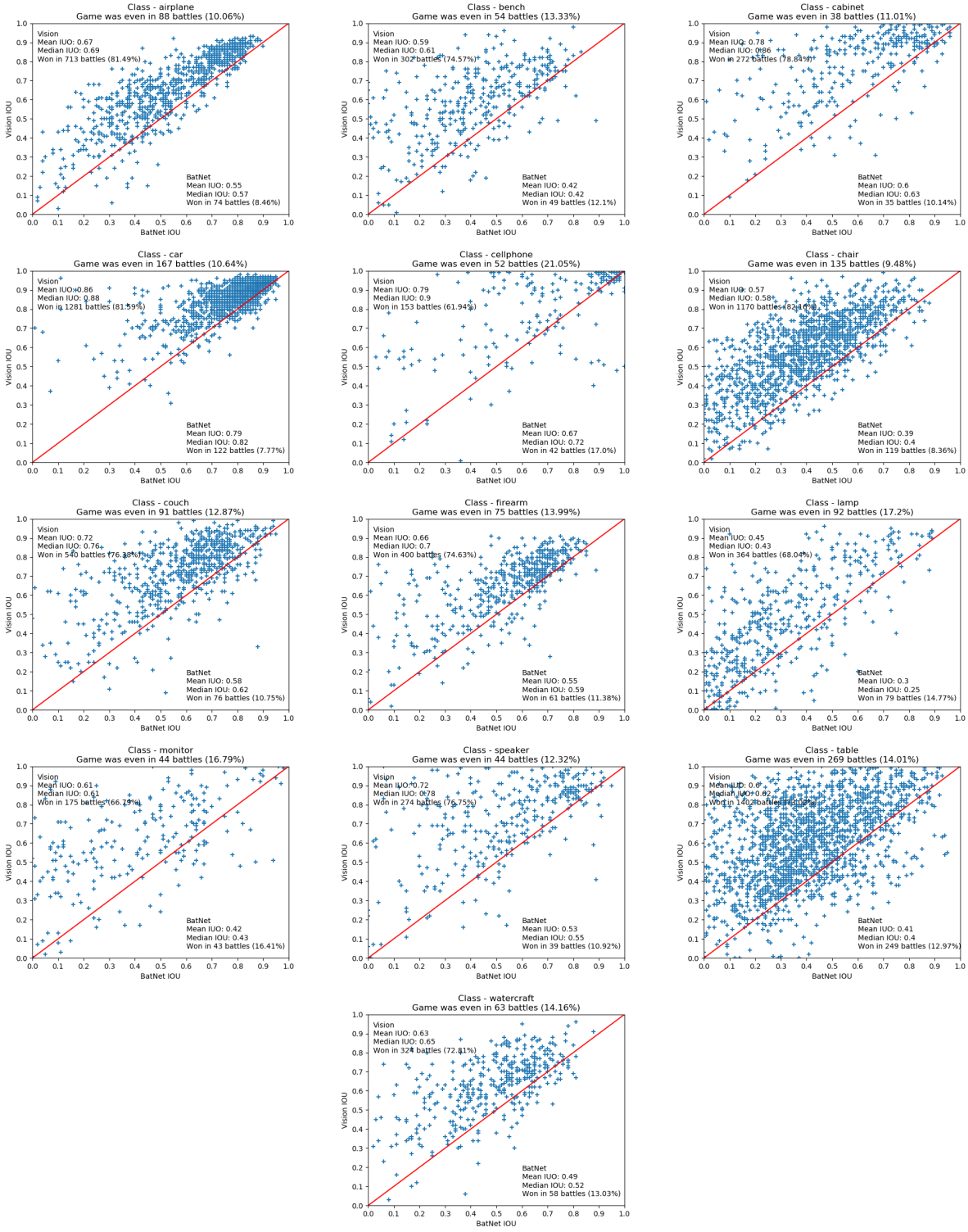

Fig. 6. IoU of Vision Vs. Sonar for the different classes

some cases, does not fall from that of vision. Furthermore, it is possible that some of the gap between the two modalities is due to the extensive research done with Vision and that as sonar and audio research evolve, this gap would be eliminated. Second, this can be done with a single echo (and two ears). Third, the 3D reconstruction can be used for echo-based object classification, showing how bats could perceive the world.

As future work, we would attempt to incorporate into the sonar, modeling mechanisms that are known to play a part in bat sensing, such as the spatial filtering of the external ear. We expect such mechanisms to further improve the reconstruction results.

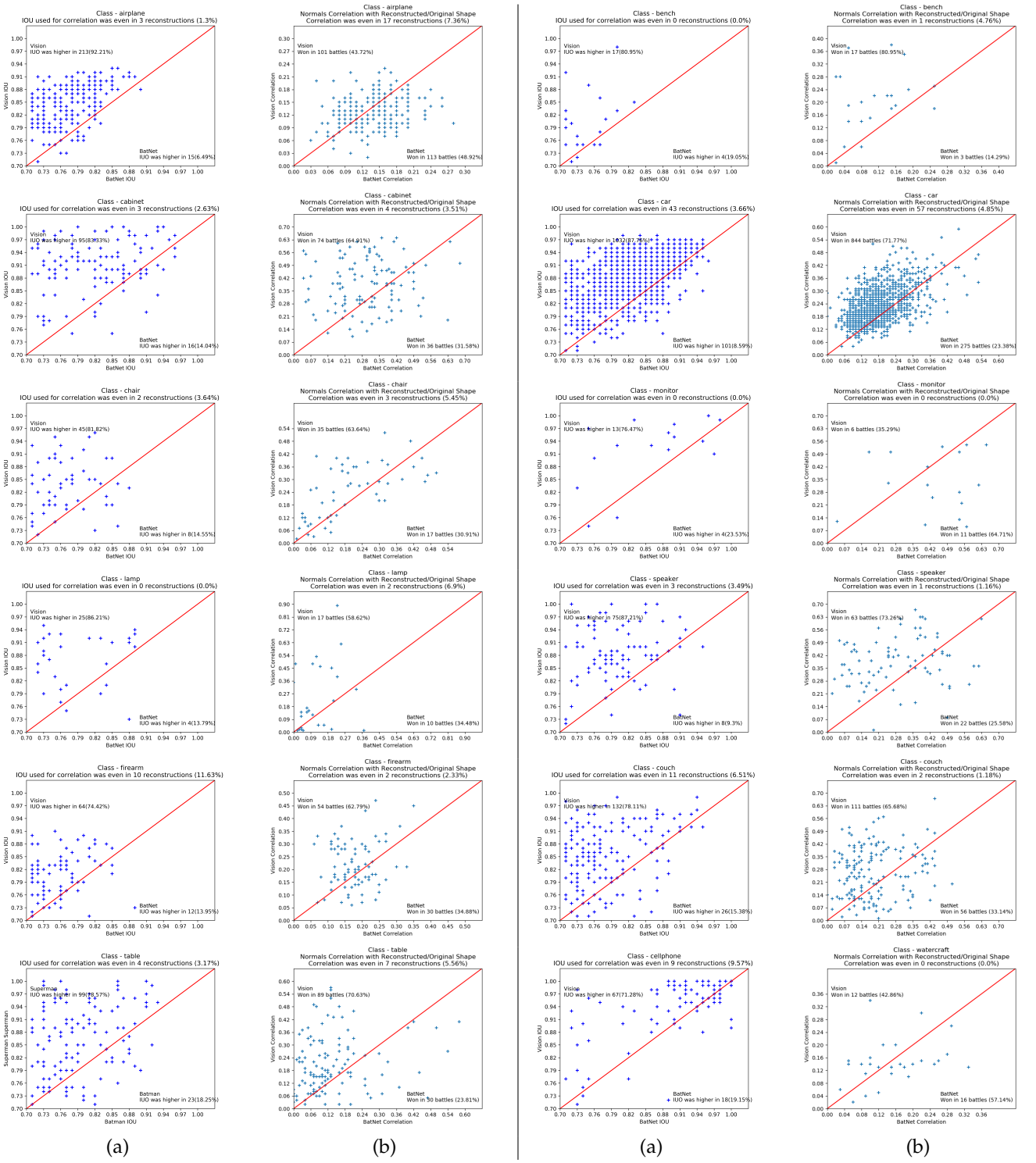

Fig. 7. IOU (a) and normal correlation score (b) for models with IOU > 0.7.

ACKNOWLEDGEMENTS

We thank Anthony Weiss for the advice and guidance. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant ERC CoG 725974).

REFERENCES

[1] G. Littwin and L. Wolf, "Deep meta functionals for shape representation," in *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[2] Y. Yovel, M. O. Franz, P. Stilz, and H.-U. Schnitzler, "Complex echo classification by echo-locating bats: a review," *Journal of Comparative Physiology A*, vol. 197, no. 5, pp. 475–490, 2011.

[3] L. Kleeman and R. Kuc, "Mobile robot sonar for target localization and classification," *The International Journal of Robotics Research*, vol. 14, no. 4, pp. 295–318, 1995.

[4] R. Müller and R. Kuc, "Foliage echoes: a probe into the ecological acoustics of bat echolocation," *The Journal of the Acoustical Society of America*, vol. 108, no. 2, pp. 836–845, 2000.

[5] Y. Yovel, P. Stilz, M. O. Franz, A. Boonman, and H.-U. Schnitzler, "What a plant sounds like: the statistics of vegetation echoes as

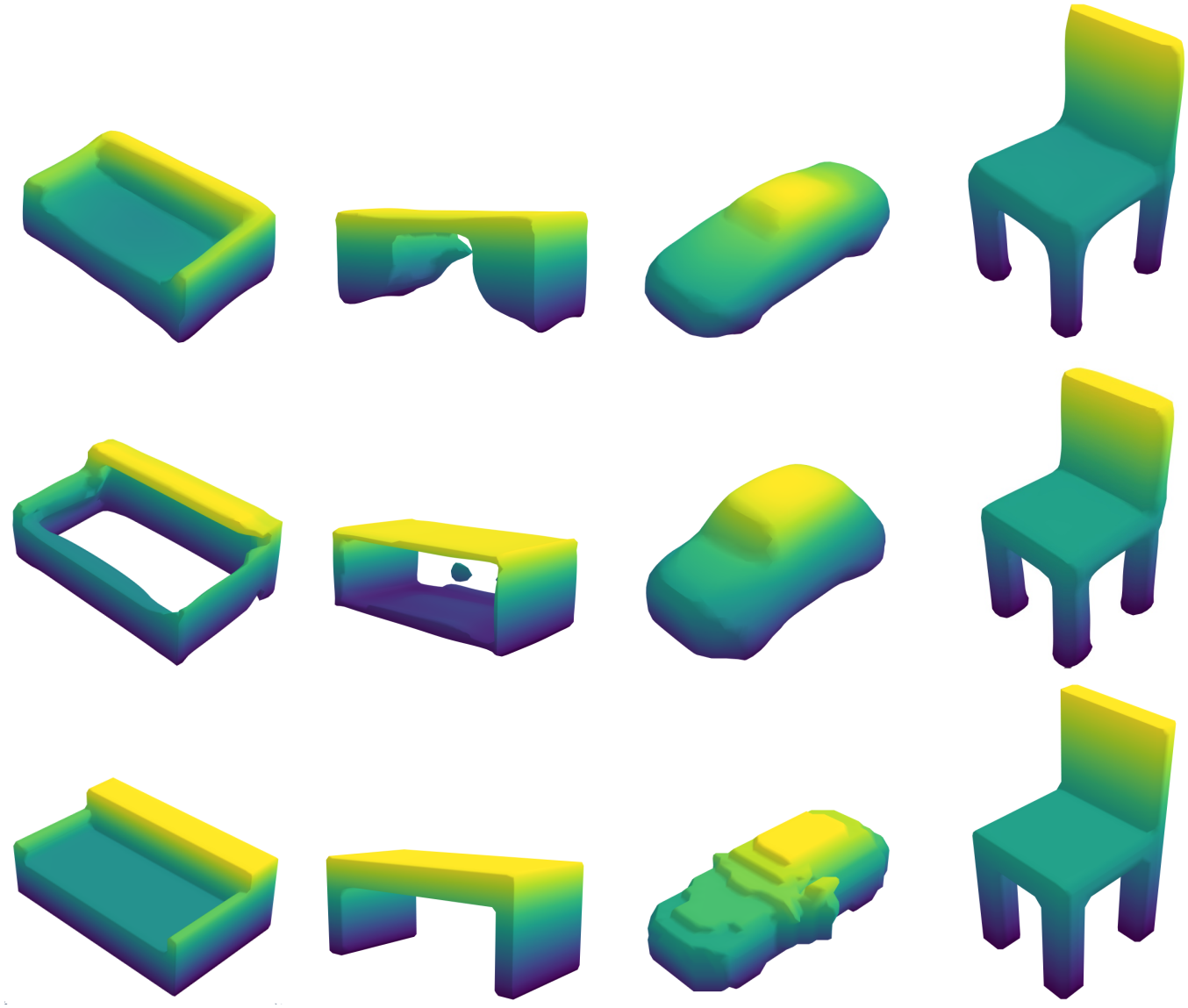

Fig. 8. Sample 3D models where the sonar-based BatNet reconstruction (top row) is better than the image-based one of [1] (middle row). The ground truth 3D models are at the bottom row.

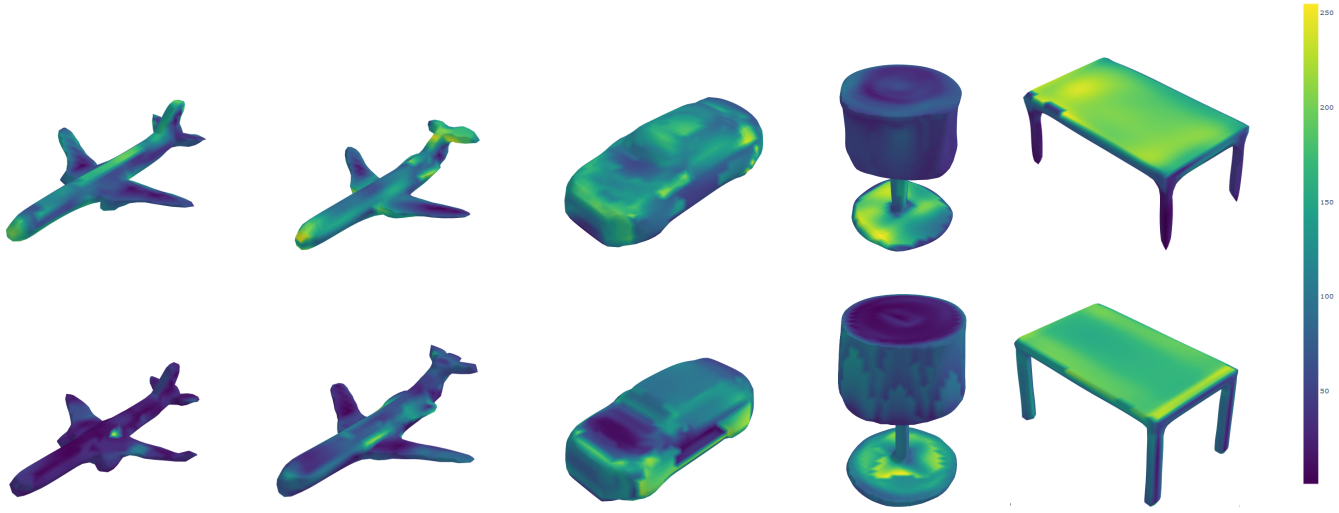

Fig. 9. Jacobian norm of (top) BatNet Vs. (bottom) Vision. color range was normalized for visualization

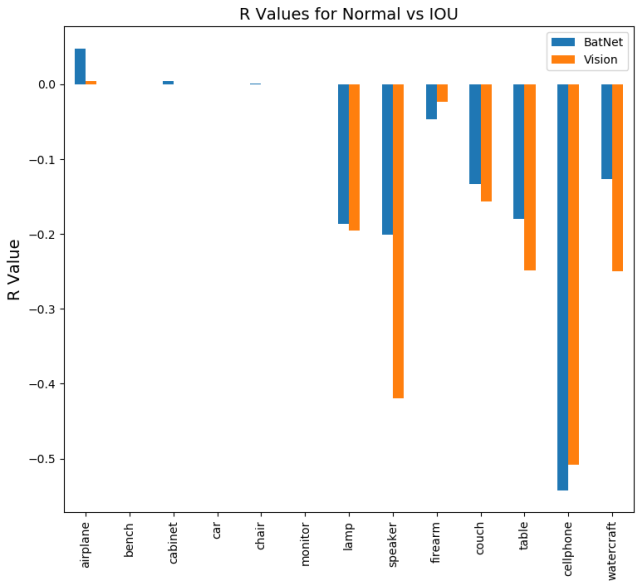

Fig. 10. Pearson correlation (R) between the variance of the angle in the ground truth 3D model, as a measure of roundness, and IOU. The computation is done on the test samples per each class for the two modalities. As can be seen, in many of the classes, round objects lead to negative correlation in both modalities.

received by echolocating bats," *PLoS computational biology*, vol. 5, no. 7, p. e1000429, 2009.

[6] C. Ming, H. Zhu, and R. Müller, "A simplified model of biosonar echoes from foliage and the properties of natural foliages," *PloS one*, vol. 12, no. 12, p. e0189824, 2017.

[7] D. Vanderelst, J. Steckel, A. Boen, H. Peremans, and M. W. Holderied, "Place recognition using batlike sonar," *Elife*, vol. 5, p. e14188, 2016.

[8] Y. Yovel, M. O. Franz, P. Stilz, and H.-U. Schnitzler, "Plant classification from bat-like echolocation signals," *PLoS Computational Biology*, vol. 4, no. 3, p. e1000032, 2008.

[9] P. K. Kroh, R. Simon, and S. J. Rupitsch, "Classification of sonar targets in air—a neural network approach," in *Multidisciplinary Digital Publishing Institute Proceedings*, vol. 2, no. 13, 2018, p. 929.

[10] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, J. Xiao, L. Yi, and F. Yu, "Shapenet: An information-rich 3d model repository," 2015.

[11] G. Riegler, A. O. Ulusoy, and A. Geiger, "Octnet: Learning deep 3d representations at high resolutions," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul 2017. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2017.701

[12] M. Tatarchenko, A. Dosovitskiy, and T. Brox, "Octree generating networks: Efficient convolutional architectures for high-resolution 3d outputs," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2088–2096.

[13] C. Häne, S. Tulsiani, and J. Malik, "Hierarchical surface prediction for 3d object reconstruction," in *2017 International Conference on 3D Vision (3DV)*. IEEE, 2017, pp. 412–420.

[14] S. R. Richter and S. Roth, "Matryoshka networks: Predicting 3d geometry via nested shape layers," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1936–1944.

[15] H. Kato, Y. Ushiku, and T. Harada, "Neural 3d mesh renderer," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3907–3916.

[16] S. Liu, W. Chen, T. Li, and H. Li, "Soft rasterizer: Differentiable rendering for unsupervised single-view mesh reconstruction," *arXiv preprint arXiv:1901.05567*, 2019.

[17] N. Wang, Y. Zhang, Z. Li, Y. Fu, W. Liu, and Y.-G. Jiang, "Pixel2mesh: Generating 3d mesh models from single rgb images," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 52–67.

[18] H. Fan, H. Su, and L. J. Guibas, "A point set generation network for 3d object reconstruction from a single image," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[19] L. Jiang, S. Shi, X. Qi, and J. Jia, "Gal: Geometric adversarial loss for single-view 3d-object reconstruction," in *The European Conference on Computer Vision (ECCV)*, September 2018.

[20] Z. Chen and H. Zhang, "Learning implicit fields for generative shape modeling," 2018.

[21] J. J. Park, P. Florence, J. Straub, R. Newcombe, and S. Lovegrove, "Deepsdf: Learning continuous signed distance functions for shape representation," 2019.

[22] L. Mescheder, M. Oechsle, M. Niemeyer, S. Nowozin, and A. Geiger, "Occupancy networks: Learning 3d reconstruction in function space," 2018.

[23] B. Klein, L. Wolf, and Y. Afek, "A dynamic convolutional layer for short range weather prediction," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 4840–4848.

[24] G. Riegler, S. Schulter, M. Rüther, and H. Bischof, "Conditioned regression models for non-blind single image super-resolution," in *2015 IEEE International Conference on Computer Vision (ICCV)*, Dec 2015, pp. 522–530.

[25] X. Jia, B. De Brabandere, T. Tuytelaars, and L. V. Gool, "Dynamic filter networks," in *Advances in Neural Information Processing Systems*, 2016, pp. 667–675.

[26] D. Ha, A. Dai, and Q. V. Le, "Hypernetworks," *arXiv preprint arXiv:1609.09106*, 2016.

[27] L. Bertinetto, J. F. Henriques, J. Valmadre, P. Torr, and A. Vedaldi, "Learning feed-forward one-shot learners," in *Advances in Neural Information Processing Systems*, 2016, pp. 523–531.

[28] J. Huang, H. Su, and L. Guibas, "Robust watertight manifold surface generation method for shapenet models," *arXiv preprint arXiv:1802.01698*, 2018.

[29] S. Kirkup, *The Boundary Element Method in Acoustics*, 01 2007, vol. 8.

[30] A. Boonman, B. Fenton, and Y. Yovel, "The benefits of insect-swarm hunting to echolocating bats, and its influence on the evolution of bat echolocation signals," *PLoS computational biology*, vol. 15, no. 12, p. e1006873, 2019.

[31] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (elus)," *arXiv preprint arXiv:1511.07289*, 2015.

[32] Y. Aytar, C. Vondrick, and A. Torralba, "Soundnet: Learning sound representations from unlabeled video," in *Advances in Neural Information Processing Systems*, 2016.

[33] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016, pp. 770–778.

[34] K. Choi, G. Fazekas, M. Sandler, and K. Cho, "Convolutional recurrent neural networks for music classification," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 2392–2396.

[35] S. Adavanne, A. Politis, J. Nikunen, and T. Virtanen, "Sound event localization and detection of overlapping sources using convolutional recurrent neural networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. PP, pp. 1–1, 12 2018.

[36] C. B. Choy, D. Xu, J. Gwak, K. Chen, and S. Savarese, "3d-r2n2: A unified approach for single and multi-view 3d object reconstruction," in *European conference on computer vision*. Springer, 2016, pp. 628–644.

[37] L. van der Maaten and G. Hinton, "Visualizing high-dimensional data using t-sne," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.