

Next-best-view Regression using a 3D Convolutional Neural Network

J. Irving Vasquez-Gomez · David Troncoso · Israel Becerra · Enrique Sucar · Rafael Murrieta-Cid

Received: date / Accepted: date

Abstract Automated three-dimensional (3D) object reconstruction is the task of building a geometric representation of a physical object by means of sensing its surface. Even though new single view reconstruction techniques can predict the surface, they lead to incomplete models, specially, for non commons objects such as antique objects or art sculptures. Therefore, to achieve the task's goals, it is essential to automatically determine the locations where the sensor will be placed so that the surface will be completely observed. This problem is known as the next-best-view problem. In this paper, we propose a data-driven approach to address the problem. The proposed approach trains a 3D convolutional neural network (3D CNN) with previous reconstructions in order to regress the position of the next-best-view. To the best of our knowledge, this is one of the first works that directly infers the next-best-

view in a continuous space using a data-driven approach for the 3D object reconstruction task. We have validated the proposed approach making use of two groups of experiments. In the first group, several variants of the proposed architecture are analyzed. Predicted next-best-views were observed to be closely positioned to the ground truth. In the second group of experiments, the proposed approach is requested to reconstruct several unseen objects, namely, objects not considered by the 3D CNN during training nor validation. Coverage percentages of up to 90 % were observed. With respect to current state-of-the-art methods, the proposed approach improves the performance of previous next-best-view classification approaches and it is quite fast in running time (3 frames per second), given that it does not compute the expensive ray tracing required by previous information metrics.

This work was partially supported by CONACYT-cátedra 1507 project.

J. Irving Vasquez-Gomez
Consejo Nacional de Ciencia y Tecnología (CONACYT) - Instituto Politécnico Nacional
E-mail: jvasquezg@ipn.mx

David Troncoso
Consejo Nacional de Ciencia y Tecnología (CONACYT) - Escuela Superior de Cómputo (ESCOM), Instituto Politécnico Nacional (IPN), México City, México.

Israel Becerra
Consejo Nacional de Ciencia y Tecnología (CONACYT) - Centro de Investigación en Matemáticas CIMAT, Guanajuato, México

Enrique Sucar
Instituto Nacional de Astrofísica Óptica y Electrónica (INAOE), Puebla, México.

Rafael Murrieta-Cid
Centro de Investigación en Matemáticas CIMAT, Guanajuato, México

Keywords Object reconstruction · 3d modeling · range sensing · next-best-view · deep learning

1 Introduction

Automated three-dimensional (3D) object reconstruction or inspection is the process of building a 3D representation of a physical object by means of sensing its surface [31]; its recent applications include inspection of airplanes [13] or the reconstruction of heritage sites [34]. Due to the limited field of view of current sensors and incomplete models generated by single view reconstructions (in particular for non common objects like antique objects or art sculptures), the 3D models need to be completed by placing a visual or range sensor at several locations while the information is integrated into a partial model.

While state-of-the-art techniques for surface sensing and model integration are already mature enough for the task, for example, in Visual Simultaneous Localization and Mapping techniques [28] [22], the search of the optimal sensing locations remains as an open problem with growing interest by the robotics and computer vision community [4].

Early work has defined the aforementioned problem as the computation of the next-best-views (NBV) [5], where each NBV is the sensor view (position and orientation) that maximizes the reconstructed surface while the positioning and registration constraints are satisfied [31]. Current methods can be classified into search-based or surface-based methods. In search-based methods, a set of candidate views is generated and then evaluated by a utility function [6][37] [7]. On the other hand, in surface-based methods, the reconstructed surface is analyzed to determine the NBV [26][3]. Such methods require large computation times (search-based) or they are limited by object’s auto occlusions (surface-based). Recent paradigms, in [25] and [43], have addressed the problem of NBV planning as a supervised learning problem where the objective is to find a function that predicts the NBV using previous knowledge. In the case of [25], we have proposed a method for generating datasets and a 3D convolutional neural network (3D-CNN), called NBV-Net, for predicting the position of the NBV. However, such previous methods are restricted to a classification, namely, the output of the 3D-CNN is limited to a small set of possible sensor views. In a real reconstruction case, such limitation could lead to an incomplete model. Therefore, it is necessary to obtain the NBV in a continuous domain. Unlike previous approaches, this paper presents further progress that is summarized as follows:

1. The works in [25] and [43] address a classification problem, while the current paper solves a regression problem. The main implication is that in this work, one is not limited to discrete predefined sensing locations; instead, the NBV is determined in the continuum.
2. The present paper provides an analysis in the terms of the number of layers in the network architecture. Such an analysis is not present in [25].
3. This work also presents an analysis in terms of the presence of dropout.
4. A qualitative and quantitative analysis of predicted NBVs compared with their ground truth is provided.
5. We tested the method with thirteen new objects included neither in the training dataset [24] nor in [25].

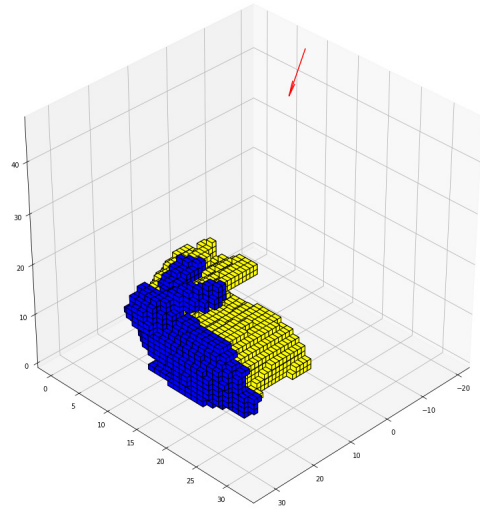


Fig. 1: Example of a predicted next-best-view. The partial model is represented by a probabilistic grid where the blue voxels indicate scanned surface and yellow voxels indicate unknown space. The predicted next-best-view is drawn in red. We can see that from its position it will observe unknown volumes while maintains an overlap with already scanned surface. Figure best seen in color.

As it was just mentioned, in this paper we propose a method for computing the NBV in a continuous domain. We have modeled the task as a regression problem where we want to learn a continuous function that receives as input the current state of the object reconstruction and predicts the NBV. To achieve the objective, we are using an extended dataset (with respect to [25]) and we are presenting a new CNN architecture. The proposed CNN regresses the position of the NBV, while the orientation is computed geometrically. We have validated the network performance using a test set and we have obtained a relatively small absolute error with respect to the ground truth NBV. In addition, the trained and validated network has been used for providing each sensor location during the reconstruction of completely unseen objects. The unseen objects are entities that were not seen by the network during training nor validation. As a result, the proposed approach has been capable of providing each sensor location in the continuous six-dimensional space (position and rotation). See Fig. 1 as an example of a computed NBV. With respect to the current approaches, the proposed method is quite fast, one third of a second, because it avoids the ray tracing step, required by the majority of current state-of-the-art methods [6][37], and it only requires a forward pass of the network.

A possible application of this work is verification of the size and shape of a given object. That is, the automatic modelling can be used to determine how well the object matches the design specifications. Another application is to use the model as an input to perform an automatic manipulation task with a robot. Moreover, the fast performance achieved in the present approach is an important step toward online NBV computation [32], in which new gathered information is used for fast NBV recalculation as the sensor moves. Decreasing the computation time of the NBV gives rise to a series of potential benefits, for instance, it allows one to execute complex tasks such as simultaneous object reconstruction and manipulation, fast building modeling and inspection which may be critical for human safety in emergency situations such as earthquakes and floods, reactive obstacle modelling and collision avoidance for reactive navigation, among others.

1.1 Comparison with related methods and paper organization

In order to show the benefits and drawbacks of the proposed method, a comparison, in terms of percentage of object reconstruction and processing time needed to compute the next-best-view, between the method proposed in this work and other two related approaches will be presented. One of the methods is based on classification [25] and the other uses information gain to exhaustively evaluate a given number of views [18].

The rest of the paper is organized as follows. Section 2 provides a brief overview of relevant recent work on the field. Section 3 provides the background of the proposed method, including the description of the used dataset. Section 4 presents the proposed method for addressing the problem. Section 5 describes the experiments, including networks training as well as the reconstruction of unseen objects. Finally, section 6 gives the conclusions and future research directions.

2 Related work

As we have already mentioned, the present paper is about the computation of the next-best-view (NBV) for 3D object reconstruction. There is a lot of work available in this area. In [31, 4, 42] interesting surveys about object reconstruction are presented. Some relevant works among many contributions are the following. The work in [36] presents a method for planning a next-best-view for object reconstruction in the workspace. The approach uses inverse kinematics computation to

get a configuration matching the desired sensor location. In [19], the authors have as a main objective to obtain a high quality surface model allowing applications such as grasping and manipulation. That work integrates 3D modeling, autonomous view planning and motion planning in a coherent manner.

The authors in [19] use Probabilistic Road Maps (PRMs) [15] and Rapidly-Exploring Random Trees (RRTs) [21] to find collision free paths. In [16], the authors propose a method to determine the next-best-view for an efficient reconstruction of highly accurate 3D models. The approach is based on the classification of the acquired surfaces, it also combines that classification with a best view selection algorithm based on mean shift. In [29], the authors present an information gain-based variant of NBV problem for a cluttered environment. They propose a belief model that allows one to obtain an accurate prediction of the potential information gain of new viewing locations. In [6] the authors investigate which formulation of information gain is best for a volumetric 3D reconstruction with a robot, which is equipped with a dense depth sensor. The authors also provide a comparative study about the performance of information gain metrics for active 3D object reconstruction.

In [37], the authors propose a method for next-best-view/state planning for 3D object reconstruction. The proposed method generates a set of candidates in the state space, later only a subset of these views is kept by filtering the original set. A utility function that integrates several aspects of the problem and an efficient strategy to evaluate the candidate views is proposed. The work in [20] addresses NBV planning for multiple depth cameras and propose a utility function that scores sets of view-points and avoids overlap between multiple sensors. The authors proved that multi-sensor NBV planning with such utility function is an instance of submodular maximization under a matroid constraint, allowing them to propose a polynomial-time greedy algorithm.

In [32], a method is proposed for inspecting a partially known environment. First, a target goal is computed and then it determines a path until a local area is inspected. The inspection is declared as complete if the percentage of unknown volume with respect to the whole unknown volume is lower than a threshold. In [33], the same authors mentioned that the completion of a volumetric map does not necessarily describe the completion of a 3D model, consequently, the model completeness is evaluated, according to the quality of the reconstructed surfaces. Also concerning exploration and inspection of 3D-environments, the authors of [9] utilize a map representation based on a Truncated Signed

Distance Function (TSDF). The TSDF data is used to identify missing parts of the model and generate a list of candidate sensor configurations, the visit of which is scheduled using an NBV planning method.

In [30] a motion planning strategy for exploration of ground-level structures is proposed. The method has two main steps, in the first step, it follows the contour of an unknown target, then the robot moves to the missing portions of the reconstructed model. The work in [27] consider the 3D model reconstruction problem in the context of infrastructure maintenance. Their approach first reconstructs an approximate 3D model using only sparse point clouds generated from a Structure-from-Motion method; the resulting rough model can be used to predict the quality of the final dense model and for an optimization-based view planning based on degraded regions.

On the other hand, there are some related methods based on machine learning. The majority address view planning for object recognition. For example, Wu et al. [40] proposed a deep belief network to perform surface prediction for model completion in order to achieve a faster object recognition. Such predicted surface is used to guide the next sensing location. Johns et al. [12] apply deep learning for computing the sequence of views that increase the mutual probability of recognizing an object. Their method uses pairs of views such that the problem become tractable. Their method uses two neural networks, one for predicting the object class and one for predicting the next-best-view from a predefined view sphere. Even-though they show promising results, the views are still limited to a given set and their proposed method can not be directly applied to object reconstruction. In consequence, for dealing with object reconstruction, Hepp et al. [10] proposes to learn a function that measures the utility of a candidate view. The supervised learning process uses known volumetric maps to determine the ground truth utility. Bai. [1] proposes an exploration method based on a deep neural network for selecting the robot 2d position from a discrete set of positions. The exploration method determines the best position in two steps: first, the network outputs some candidate positions, then, those candidates are evaluated by Mutual Information [14]. The strategy proposed by Wang et al. [38] combines a hand-crafted utility with a learned utility. The learned utility is structured as a classification approach with a CNN based on AlexNet. Inside it, the input is a range image and the outputs is a vector with the scores of a fixed set of views. In [39], a NBV method is proposed that leverages deep learning for phenotyping plants. The approach uses a network based on the Point Completion Network (PCN) [41], but with the capability to predict

the confidences of completed points. The network learns the structural prior of plants, receives cloud points that partially model the plant in question, and outputs a completed point cloud of the plant. Subsequently, the resulting point cloud is used to build a predicted octomap model, which in turn is used to guide NBV planning; the next-best viewpoint is defined as the one that can provide the maximum amount of information for the plant phenotyping. Zeng et al. [43] proposes a deep neural network for evaluating a set of candidate views efficiently. Their method, contrary to [25], require as input the reconstructed point cloud, avoiding the need of an intermediate representation such as the probabilistic grid. Zeng et al. architecture provides efficient evaluation. However, their method is still limited to a predefined set of views. Therefore, to the best of our knowledge, there is no method for estimating directly the position of the next-best-view in the 3D space for 3D object reconstruction.

3 Background

In this section, we provide the background of the proposed method for learning the next-best-view.

3.1 3D reconstruction

The automated 3D reconstruction is an iterative process of four steps: positioning, sensing, registration-and-update and next-best-view planning. The positioning places the sensor at the desired pose, then the surface is measured by the sensor, next, the registration-and-update transforms the observed surface to a common reference frame [2] and integrates the information into a single partial model [11], then the next-best-view planning determines the next sensor pose. The process is repeated until an stop condition.

Some assumptions that are made in this paper are the following: the object of interest is encapsulated in a cube, the center of the cube is assumed as the center of the object, the sensor is capable of obtaining a point cloud from the object's surface.

3.2 Partial model

The partial model, \mathcal{M} , stores the accumulated information about the object of interest. In this paper, it is implemented with a probabilistic grid where the space is evenly divided into small cubes. Each cube is called voxel and has associated a probability of being occupied. The occupancy probability is updated with a Bayes

filter using the perceptions from the sensor [35]. Considering m , n and o as the dimensions of the grid, the partial model will also be written as $\mathbb{R}^{m \times n \times o}$. A reconstruction state is an instance of the partial model; two partial models can contain the same object but they can differ in the reconstruction state. In this paper, we are using the Octomap library [11] to implement the probabilistic grid.

3.3 Next-best-view

The next-best-view is defined as a six-tuple:

$$v = (x, y, z, \alpha, \beta, \gamma) \quad (1)$$

where x , y and z define a position in the 3D euclidean space and α , β and γ define the yaw, pitch and roll orientations according to the Tait-Bryan angles. Based on the orientations, a rotation matrix, $R(\alpha, \beta, \gamma)$, that transforms the sensor pose can be directly obtained from the rotation angles.

3.4 Dataset

The dataset that we are using was proposed in a previous work from our research group [25]. This dataset contains tuples of regressors and responders. The regressor is a partial model, $\mathcal{M}_i \in \mathbb{R}^{32 \times 32 \times 32}$, and the responder is its corresponding next-best-view, $v_i = (x, y, z, \alpha, \beta, \gamma)$, where i is an index over the dataset. The dataset was build by performing several reconstructions for different objects. The reconstruction scene places the object at the origin of the global reference frame and a large set of possible sensor locations are generated forming a sphere of radius 0.4 m. 12 object shapes were included in the dataset. During the reconstructions, for each partial model a next-best-view was computed by performing an exhaustive search over the set of possible views. The view that maximized the increment of reconstructed surface and satisfy an overlap was taken as the ground truth NBV. In our previous work [25], we restrict possible predictions to a finite set of views (14 classes) around the object. Unlike it, in this paper, we are not restricting the predictions, therefore the possible predictions are in the 6D space. The dataset that we are using in this paper is available at [24] and the reduced dataset used in our previous work is available at [23].

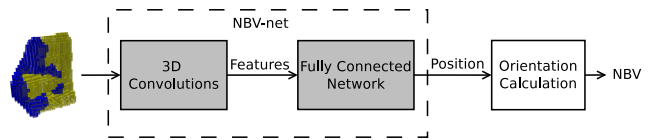


Fig. 2: Overall regression approach for next-best-view planning.

4 Next-best-view regression

The next-best-view regression is the task of finding a function

$$f(\mathcal{M}) : \mathbb{R}^{m \times n \times o} \rightarrow \mathbb{R}^3 \times SO(3) \quad (2)$$

so that, the information perceived in $\hat{v} = f(\mathcal{M})$ increases the surface of the object contained in the partial model \mathcal{M} satisfying the registration and positioning constraints [25]. Notice that the input of f is the probabilistic grid where $m \times n \times o$ is the number of voxels and the output is composed by a position in \mathbb{R}^3 and a orientation represented by a rotation matrix in the special orthogonal group $SO(3)$.

4.1 Regression approach

In the proposed scheme, we use a 3D convolutional neural network, denoted by Φ , to regress the position of the NBV. Then the orientation is computed by aligning the sensor to the center of the object. This approach has the advantage of being easier to train, given that only three continuous variables, corresponding to the position, are predicted by Φ , while the orientation is computed using geometric reasoning. Fig. 2 depicts the whole approach to compute the NBV. Formally, the network output is:

$$\hat{p} = \Phi(\mathcal{M}) \quad (3)$$

where $\hat{p} = (x_{\hat{p}}, y_{\hat{p}}, z_{\hat{p}})$ is a position in \mathbb{R}^3 where the sensor will be placed. Given that in running time the objects can have different sizes w.r.t. the training samples, the predicted position (\hat{p}) must be scaled in order to fit the object size. This is achieved by maintaining the same grid resolution ($32 \times 32 \times 32$) but changing the voxel size according to the object size and scaling the predicted position by a factor k :

$$s = k \cdot \hat{p} \quad (4)$$

The k factor can be calculated so that the sensor's field-of-view encloses the object of interest, that is, given the smallest sensor's opening angle and the object's major span, the k factor is the distance such that the object lies within such an opening angle.

Once the position is predicted by the network, the orientation is computed first as a unit vector,

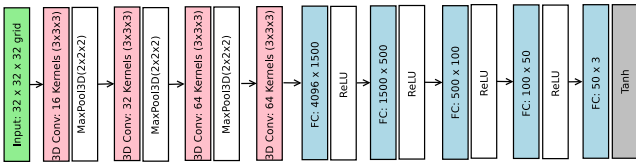


Fig. 3: NBV-net 4-5 architecture. The number 4-5 stands for 4 feature extraction layers and 5 fully connected layers.

$\hat{r} = (x_r, y_r, z_r)$, indicating the orientation of the sensor’s director ray, namely:

$$\hat{r} = \frac{c - \hat{p}}{\|c - \hat{p}\|} \quad (5)$$

where c is the center of the object. Then, \hat{r} is converted to Euler rotation angles:

$$\alpha = \arctan\left(\frac{y_r}{x_r}\right) \quad (6)$$

$$\beta = \arcsin(z_r) \quad (7)$$

It is worth to say that under this approach, \hat{r} provides the yaw and pitch parameters but the roll parameter (rotation over the camera axis) is omitted. In consequence, $\gamma = 0$. Even though one degree of freedom is lost, the roll angle is usually omitted in next-best-view planning because it has the lowest impact on the built model. Finally, the predicted NBV is given by:

$$\hat{v} = (s_x, s_y, s_z, \alpha, \beta, \gamma) \quad (8)$$

4.2 NBV-net

In a previous work [25], our research group presented a 3D CNN for view classification. However, it has shown poor performance for the regression task, as we will discuss later in the experiments. Therefore, we propose to extend our previous network to the regression task.

The proposed architecture, NBV-net 4-5, receives as input a probabilistic grid of dimension $32 \times 32 \times 32$ and predicts the position of the NBV. It has four 3D convolutional layers and five connected layers (giving the name 4-5). The 3D convolutional layers extract features from the grid to a 4D vector, then the features are flattened and passed through the set of fully connected layers. See Fig. 3.

To simplify the detailed network description, we will use the following notation: $C(f, k, s)$ defines a 3D convolutional layer with f filters of size $k \times k \times k$ and stride $s \times s \times s$, $P(s)$ a max pooling layer of stride $s \times s \times s$ and $FC(n)$ defines a fully connected layer with n nodes.

Then, NBV-net 4-5 is configured as follows: $C(16, 3, 2)$, $P(2)$, $C(32, 3, 2)$, $P(2)$, $C(64, 3, 2)$, $P(2)$, $C(64, 3, 2)$, $FC(1500)$, $FC(500)$, $FC(100)$, $FC(50)$, $FC(3)$. Note that, there is no pooling after the fourth convolutional layer and the 4096 features are pass through three fully connected layers. The activations are performed by Rectified Linear Units (ReLU) except for the last one which applies a hyperbolic tangent activation function (Tanh).

FC(500), FC(100), FC(50), FC(3). Note that, there is no pooling after the fourth convolutional layer and the 4096 features are pass through three fully connected layers. The activations are performed by Rectified Linear Units (ReLU) except for the last one which applies a hyperbolic tangent activation function (Tanh).

5 Experiments

The main reason for selecting a particular CNN architecture or a set of hyper-parameters is to obtain a good generalization on the task. In this context, a good generalization will be to provide a NBV that increases the object surface despite the variability of i) the object shape and ii) the current reconstruction state. Fig. 4 can be helpful to observe the different object shapes and reconstruction states. In that sense, the dataset [24] provides more than ten thousands of examples of reconstruction states because the objects were reconstructed several times from different initial positions. However, the shapes are limited to 12. For this reason, the experiments will focus on showing the network’s architecture and parameters that provide a better generalization.

We present two groups of experiments. The first group analyzes the training and validation of several variants of the proposed architecture. The second group of experiments test the most promising network configurations in the reconstruction of several unseen objects, these objects were not seen by the CNN during training nor validation. At the end of this section, we present a discussion about the network advantages as well as the current challenges. For all the experiments, the architectures were implemented in PyTorch. The experiments were carried out using an Intel i7 machine with NVIDIA Geforce 1080 GPU.

5.1 Network training and architecture variants

In this group of experiments, we analyze the training and validation performance of the proposed architecture. First, we compare and analyze several variants of the architecture including those reported previously. Then, we analyze the use of the regularization method dropout during training.

5.1.1 Architecture variants

Several architectures for 3D regression problems have been proposed, for example to determine the pose of a known object. However, in NBV planning for 3D object reconstruction, the number of current methods is very limited. Therefore, we compare the proposed network

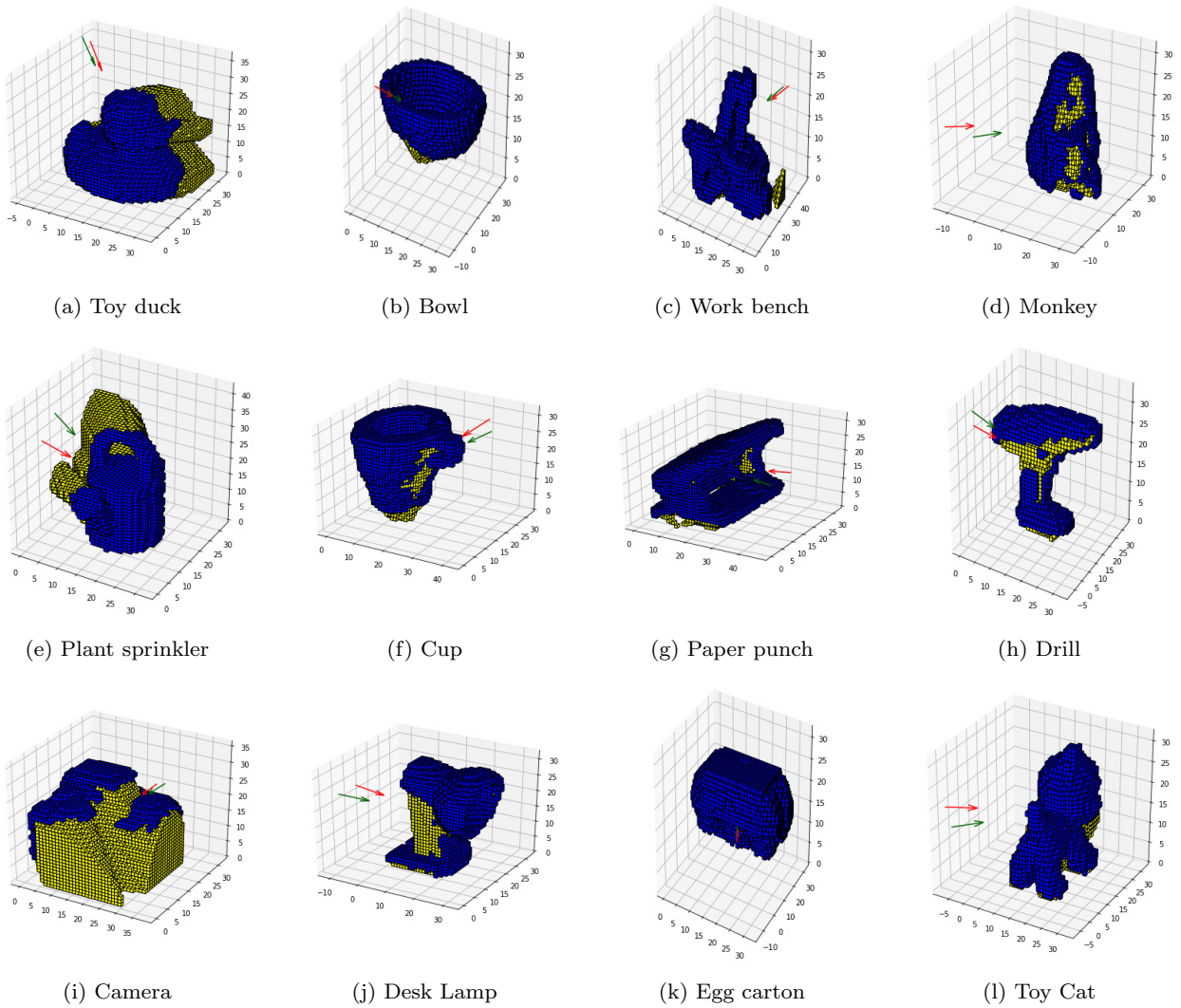


Fig. 4: Comparison of the predicted next-best-view versus the ground truth for several objects in the dataset. Blue voxels indicate measured surface. Yellow voxels indicate unknown space. The predicted next-best-view is drawn in red. The ground truth next-best-view is drawn in green. Figure best seen in color.

versus three variants, where one of them is a modified version of an architecture reported for classification.

- NBV-net 3-3. The shortest network; it includes three convolutional layers and 3 fully connected layers. In detail, $C(10,3,2)$, $P(2)$, $C(12,3,2)$, $P(2)$, $C(8,3,2)$, $P(2)$, $FC(1024)$, $FC(500)$, $FC(3)$.
- NBV-net 3-5. Network proposed in [25] for NBV classification. The last layer is replaced by three nodes with a *Tanh* function as the activation function. In detail, $C(10,3,2)$, $P(2)$, $C(12,3,2)$, $P(2)$, $C(8,3,2)$, $P(2)$, $FC(1500)$, $FC(500)$, $FC(100)$, $FC(50)$, $FC(3)$.
- NBV-net 4-3. This variant increases the number of feature extraction layers to 4 but decreases the fully connected layers to 3. In detail, $C(16,3,2)$, $P(2)$,

$C(32,3,2)$, $P(2)$, $C(64,3,2)$, $P(2)$, $C(64,3,2)$, $FC(1024)$, $FC(500)$, $FC(3)$.

- NBV-net 4-5. Both, feature extraction and fully connected layers are increased. Description presented in section 4.2.

5.1.2 Training

According to the proposed model, we require to predict the NBV position, equation (3). Thus, the ground truth positions are normalized to unit vectors. Then, the networks are trained to reduce the Mean Squared Error

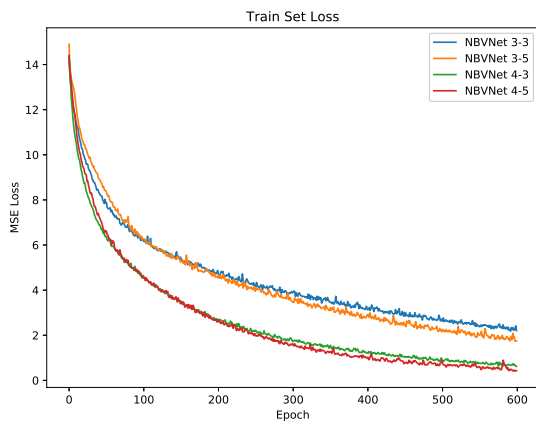


Fig. 5: Training loss.

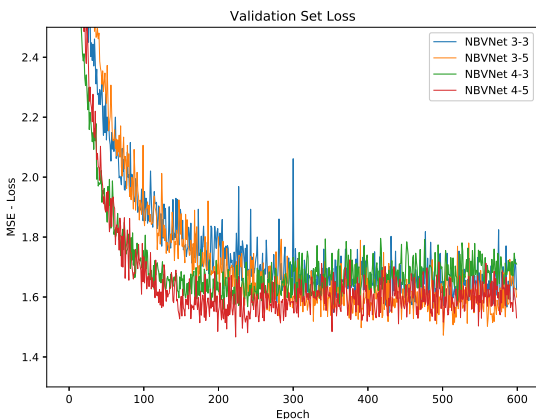


Fig. 6: Validation loss.

(MSE) loss between prediction \hat{p} and ground truth position p :

$$\text{loss}(p, \hat{p}) = \frac{1}{N} \sum_{i=1}^N (\hat{p}_i - p_i)^2 \quad (9)$$

where $N = 3$ is the number of elements in the position vector. To minimize the error, we use the Adam optimizer [17] which auto-adjusts the learning rate.

The dataset was divided randomly into 80% for training and 20% for validation. In this way, both subsets contain examples from the 12 objects but with different reconstruction states, this imply a different distribution of the occupied, unknown and free voxels. The training was performed during 600 epochs with learning rate 0.0001 and batch size 250. Each network training required an average time of five hours.

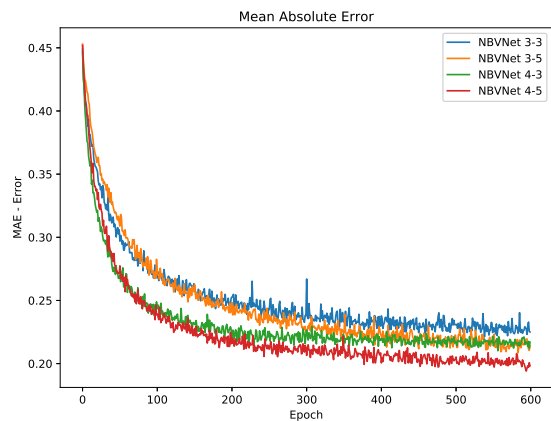


Fig. 7: Mean absolute error (MAE) over validation set. The graph shows how far in the euclidean space are the predictions from the ground truth next-best-view.

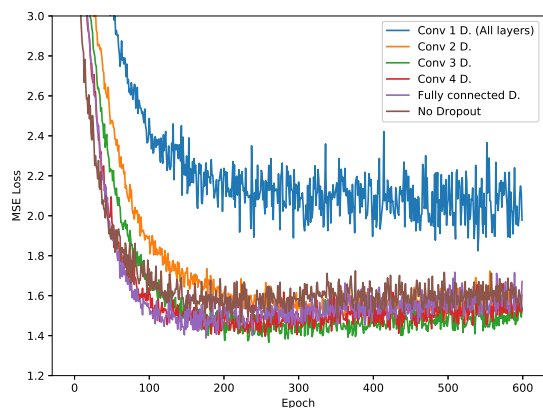


Fig. 8: Mean square error (MSE) over validation set for different dropout configurations. The regularization method dropout is inserted at several network layers. Nomenclature “Conv N D” indicates that dropout is included from convolutional layer N until the last network layer. Network tested: NBV-net 4-5.

5.1.3 Precision analysis

The training loss, calculated with eq. (9), for the training set is shown in Fig. 5. As we can see, all networks reduce the loss, but the best fitting ones, according to MSE, are the networks that include four convolutional layers instead of only three. On the other hand, with respect to the validation set, all networks reach a similar loss, see Fig. 6; however, we can observe that the networks with four convolutional layers start to overfit after 200 epochs, namely, the loss increases for such

networks. For that reason, we stop growing the depth of the proposed network. Some examples of the predictions made by NBV-net 4-5 are shown in Fig. 4.

Previous analysis focus on training performance, however, in order to provide additional information about how well in the Euclidean space the predictions are, Fig. 7 shows the mean absolute error (MAE) as the training epochs advance. Note that MAE was not used for training.

$$MAE(p, \hat{p}) = \frac{1}{N} \sum_i^N |\hat{p}_i - p_i| \quad (10)$$

We can observe from Fig. 7, that NBV-net 4-5 was the one with the best result, in terms of the smallest Euclidean distance between the predictions and the ground truth next-best-view. NBV-net 4-5 is the architecture that has the largest number of convolutional and fully connected layers. We can also observe that NBV-net 4-3 has also a good performance, particularly in the first epochs, where it has better results than NBV-net 4-5.

5.1.4 Regularization

Regularization is used for avoiding overfitting over the training set. In this context, we want to prevent learning the object’s shapes in the dataset. Therefore, in this experiment we analyze how the regularization dropout method affects network performance. The experiment inserts dropout with probability 0.5 layer by layer, starting from the fully connected layer until the first convolutional. Fig. 8 shows the MSE over the validation set using the NBV-net 4-5. As we can see in the graph, after 600 epochs, inserting dropout from the third convolutional layer until the last one (Conv 3 D) leads to the smallest MSE loss, meaning that even though some intermediate features are not present the output is adequate. However, if dropout is applied from the first convolutional layer (Conv 1 D.), then the MSE increases dramatically. Our hypothesis for this phenomenon is that low level feature extraction is very important, therefore by missing any of such features the NBV regression is affected.

5.2 3D Reconstruction of unknown objects

In this experiment, we will use the the full proposed approach (Fig. 2) as the view planner in a simulated 3D reconstruction task, where an unknown object is placed in the scene and the approach has to provide each NBV until the stop criteria is reached. Unlike previous experiment, in this case there is no ground truth NBV since

the objects are unknown and they were not used during training, therefore, we will measure the capacity of the method for completing the object.

The reconstruction scene places the object over a table, except for the chair. The sensor is simulated with a range camera. The positioning systems is simulated and directly places the sensor in the planned pose. The performance metric is the percentage of coverage, calculated as the ratio between the matching scanned points and the points in the reference model given a distance of 0.001m. Scale factor k was set to 2.5. The objects to be modeled are thirteen: a sphere, the mask of Tutankhamun, a Corinthian helmet, a Egyptian sarcophagus, the Stanford bunny, the Stanford dragon, a teapot, a caster wheel, a Moai head, a butterfly valve, an armadillo, a chair and a hammer. The objects are depicted in Fig. 9. Our simulation was implemented using the view planning library (VPL) [37] and Blensor simulator [8].

5.2.1 Generalization

First, we test the different network variations on three unseen objects that were not used during training nor validation; this was done to show which architecture provides the best generalization despite training and validation performance. For this purpose, we have selected the sphere, the bunny and the dragon, which contrast to dataset objects because of either their convex or elongated shapes. For all the variants, the initial sensor location is placed in front of the object. Once the first scan is integrated to the probabilistic grid, each network makes its prediction and the reconstruction continues, in consequence, each variation follows a different sequence of sensing locations. The experiment was done for ten scans in order to observe the coverage reached for each variant.

As we can see in Figs. 10, 11 and 12, the architecture that has the largest number of convolutional and fully connected layers (NBV-net 4-5) with dropout from convolutional layer 3, was the one with the best performance, that is, the largest final reconstruction percentage. Its reconstruction coverage ranged from around 85% up to 95%. We believe that this result is due to the increment in the network’s feature extraction layers (four wider convolutional layers), also due to the dropout that avoided overfitting over the training set, and due to the five fully connected layers that enhanced the network’s capabilities to perform regression. It is worth to say that network NBV-net 3-5 with dropout, had a tendency to cover greater portions of the objects at the beginning of the scanning. A possible explanation for this phenomenon is that with fewer convolutional

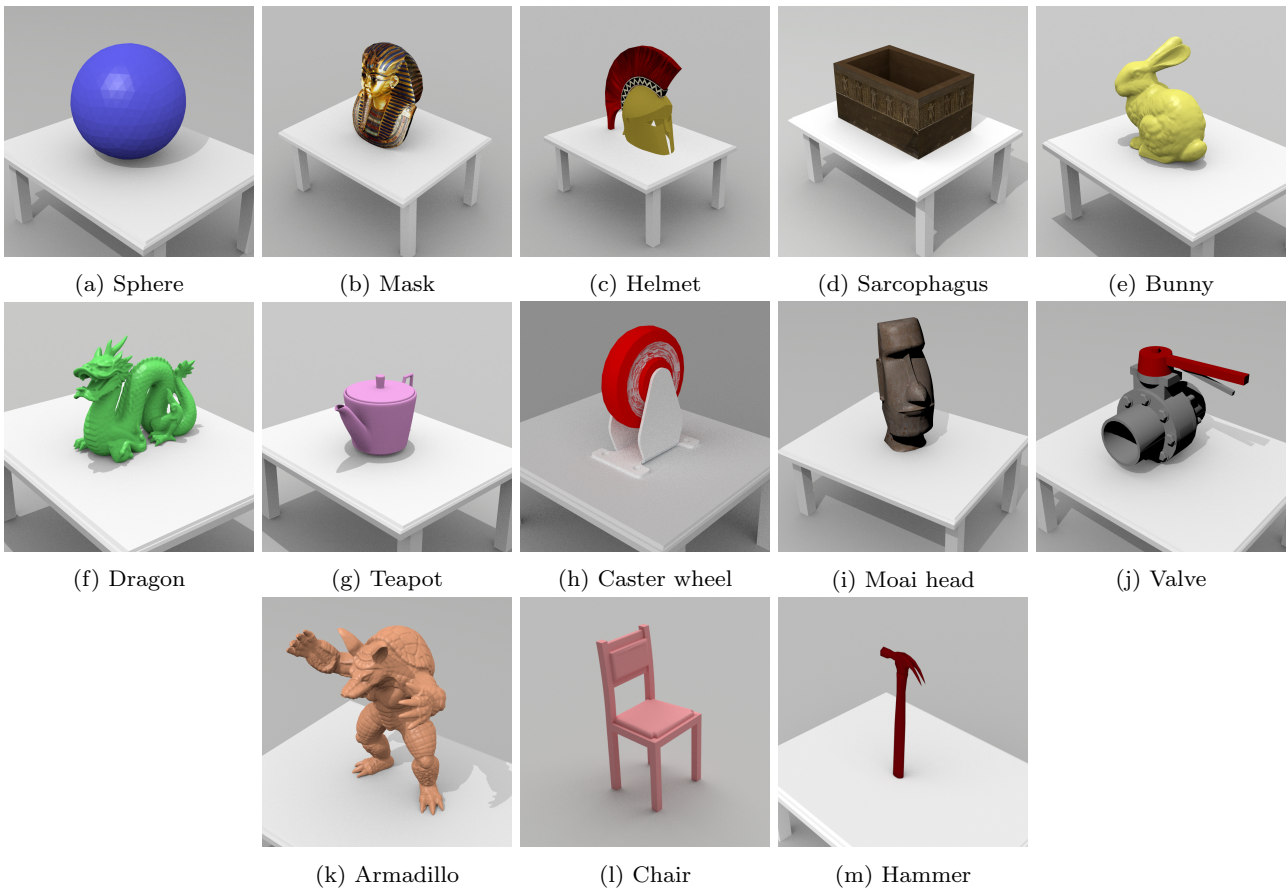


Fig. 9: 3D models used for testing the 3D CNN.

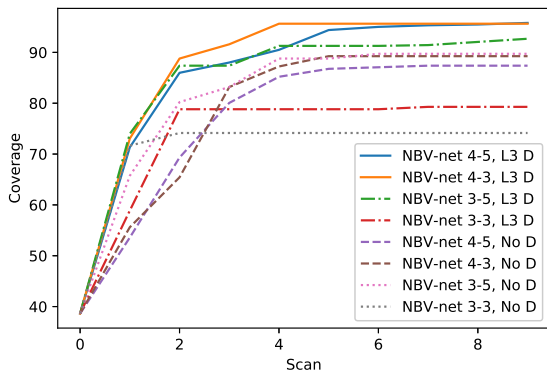


Fig. 10: Reconstruction coverage for the sphere object.

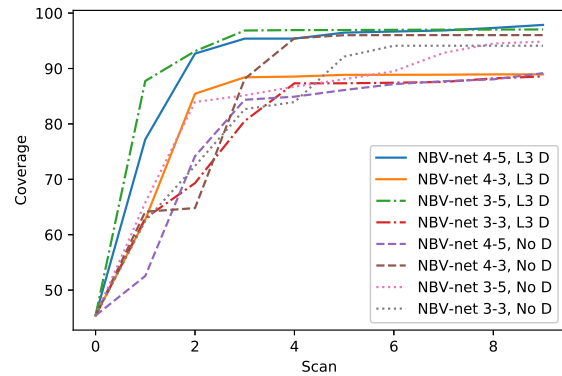


Fig. 11: Reconstruction coverage for the bunny object.

layers less features were modeled, but those were important to identify poses that provide coverage of large portions of the object; however, once the scanning progresses, the missing modeled features are important to cover details of the object, what NBV-net 4-5 was indeed able to do.

5.2.2 Different object shapes

We carry out a comparison, in terms of percentage of object reconstruction and processing time needed to compute the next-best-view, between the method proposed in this work (NBV-net 4-5) with other two related methods. In [25], the authors address the next-best-

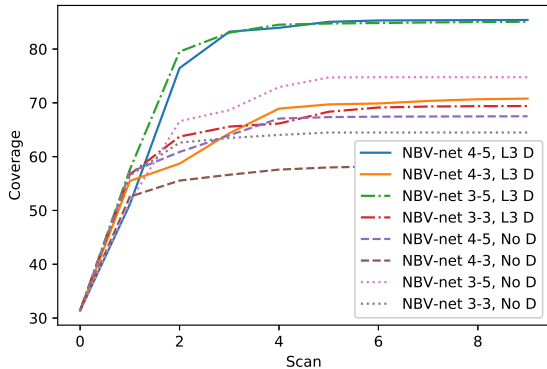


Fig. 12: Reconstruction coverage for the dragon object.

Table 1: Reconstruction coverage for each tested object.

Object	Classif.	Regression	Inf. Gain
Sphere	96.7	95.7	96.2
Mask	94.73	95.32	95.0
Helmet	82.7	84.9	86.5
Sarcophagus	64.2	71.7	94.1
Bunny	90.0	97.8	98.1
Dragon	71.3	85.4	90.4
Teapot	87.1	93.0	93.2
Caster	89.1	90.6	100
Moai Head	96.8	97.1	98.9
Valve	73.3	70.7	85.9
Armadillo	84.6	86.0	95.2
Chair	85.2	84.6	89.4
Hammer	53.2	56.8	57.0

Table 2: Processing time for next-best-view computation

	Classif.	Regression	Inf. Gain
Time	0.01 s	0.3 s	29.9 s

view problem with a classification-based approach. The output of the 3D-CNN is limited to a set of 14 possible sensor views. The approach in [18] proposed an information gain based method that exhaustively evaluates 20 views around the object. The three approaches were tested in the reconstruction of thirteen proposed unknown objects (Fig. 9). The stop criteria was ten scans.

Table 1 presents the percentage of object reconstruction and Table 2 shows the processing time needed to compute the next-best-view. the method proposed in this work is called Regression, the one proposed in [25] is labeled as Classif. and the one proposed in [18] based on information gain is labeled Inf. Gain. Even though less coverage is achieved with the proposed method compared to exhaustive search methods, the resulting models can be good enough for several tasks in which a fast decision is required based on the constructed

model. Nonetheless, this can be alleviated by adding a focused exhaustive search stage already available in the literature.

We underline that the data reported in Table 2 called processing times, are the times that the network needs to determine the next-best-view, and not the time they need to learn from examples, that is, the time the network needs to make the inference. We would also like to point out that the reported processing time for regression corresponds to architecture NBV-net 4-5, which is the one with the most layers among the tested regression architectures, and the one that takes the longest to compute the NBV. Therefore, that reported time serves as an upper bound, namely, all the tested regression architectures are able to perform the inference with a frequency of at least 3 Hz.

We can observe that the method proposed in this work, gets a larger percentage of object reconstruction for ten of the thirteen objects compared with the Classif. method. In consequence, the Regression method improves the coverage reached by the previous Classification method. On the other hand, the Inf. Gain method achieves a larger percentage of object reconstruction in eleven of the thirteen objects w.r.t. the other two methods. As a result, the exhaustive search provided by Inf. Gain. reaches the highest coverage. However, the Inf. Gain. method is two orders of magnitude slower than the one proposed in this paper. Thus, the main drawback of the Inf. Gain. approach is the large processing time that the method needs to compute the next-best-view. One can also notice that the faster method is the Classification method, which is an order or magnitude faster than the one proposed in this work. It only requires around 10 milliseconds to compute the next-best-view. Nonetheless, the main limitation of the Classification approach is that the fixed number of sensor views could lead to an incomplete model. Also note from the hammer statistics that thin objects are particularly hard to reconstruct for the three methods due to the resolution of the probabilistic grid. Fig. 13 displays the occupancy grids after reconstruction using the Regression method. Fig. 14 shows an example of the first three sensing views for the Bunny object.

5.3 Discussion

We have found that the proposed method is capable of reconstructing the majority of the object surfaces despite the object shape. The architecture that has the largest number of convolutional and fully connected layers (NBV-net 4-5), was the one with the best result, that is, largest final reconstruction percentage. The main advantage of the proposed method is its rapid

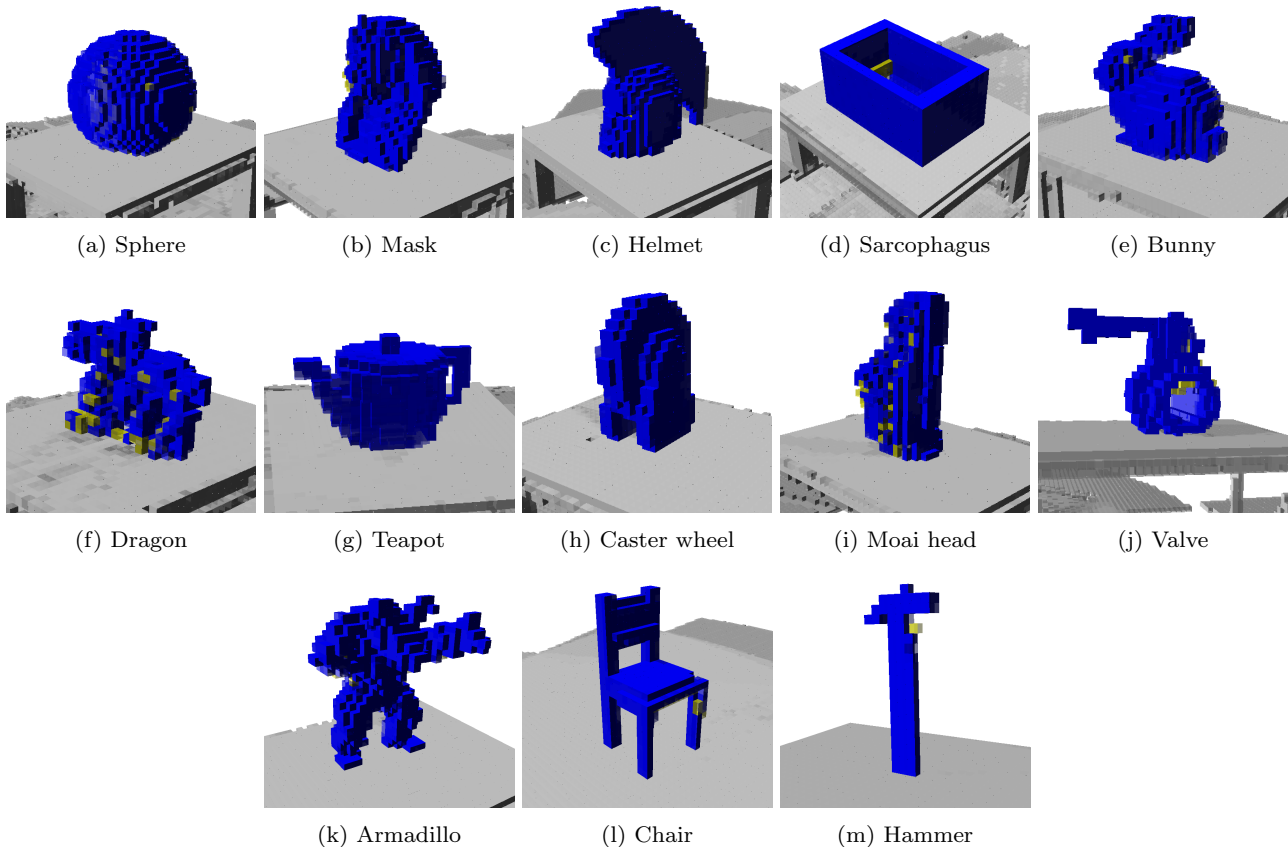


Fig. 13: Probabilistic grids after reconstruction.

response in running time (0.3 seconds). Compared with traditional next-best-view methods, the proposed one eliminates the expensive ray tracing step required to compute several information metrics. Even though training the network might be a time consuming task, it is performed only once and offline. One disadvantage could be that in some cases, it does not reach the highest coverage obtained by search-based methods, for those cases, the current method could be complemented by including a local search or a surface filling.

About the comparison between the method proposed in this work with other two methods. One is a classification-based method [25] and the other is an exhaustive search using an information gain evaluation [18]. The method based on exhaustive search and information gain gets a larger percentage of object reconstruction, but it is slow. In contrast, the method based on classification is very fast but it gets poor results in terms of percentage of object reconstruction. Based on the data in Tables 1 and 2, we conclude that the method proposed in this work gets a good tradeoff between percentage of object reconstruction and the processing time needed to compute the next-best-view. It overcomes the method

in [25] in terms of percentage of object reconstruction, and it needs a reasonable processing time to compute the next-best-view, it only takes less than a half of a second to get it.

With respect to the network training, we believe that the current approach can be improved with a new loss function that does not consider a single NBV as the ground truth. Because in several cases there is more than one good view, and they are separated in distance. However, this is not a trivial task. In addition, for future datasets it will be good to include more object shapes, leaving for the validation set different shapes (not included in the training set) as well as different reconstruction states.

6 Conclusions

We have presented a deep learning based approach for next-best-view regression. In this approach, we are addressing the next-best-view prediction in a continuous space. The proposed network architecture is designed for the particular problem and it has been trained and

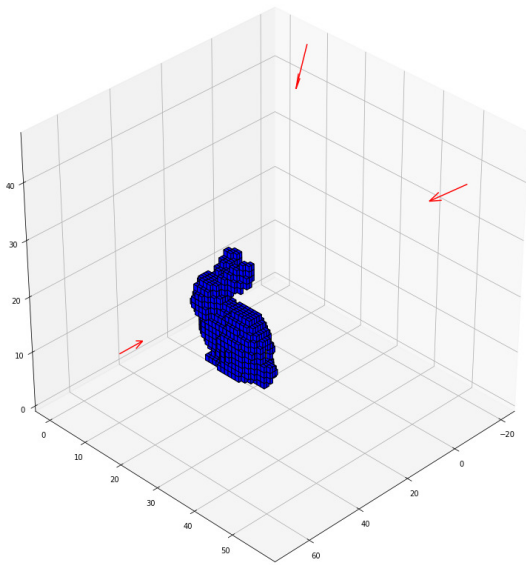


Fig. 14: First three sensing locations during the reconstruction of the Bunny object. The arrow in front of the Bunny is the initial view. The remaining arrows were predicted by the Regression method.

validated. Our experiments have shown that the proposed method generalizes well to object shapes that have not been seen by the network during training nor validation. The fast response of the proposed method is one of its advantages given that it eliminates the expensive ray tracing required by state of the art methods. We have presented a comparison between the method proposed in this work, with other two related approaches. We can conclude that the method proposed in this work gets a good trade-off between percentage of object reconstruction and the processing time needed to compute the next-best-view. For future research, we will study new loss functions as well as applications to the reconstruction of large scale buildings. Finally, it is planned to continue expanding the training and validation datasets including additional objects.

References

- Bai, S., Chen, F., Englot, B.: Toward autonomous mapping and exploration for mobile robots through deep supervised learning. In: 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 2379–2384. IEEE (2017)
- Besl, P., McKay, N.: A method for registration of 3-d shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **14**, 239–256 (1992). DOI 10.1109/34.121791
- Chen, S., Li, Y.: Vision sensor planning for 3-d model acquisition. *IEEE Transactions on Systems, Man, and Cybernetics* **35**(5), 894–904 (2005)
- Chen, S., Li, Y., Kwok, N.M.: Active vision in robotic systems: A survey of recent developments. *International Journal of Robotics Research* **30**(11), 1343–1377 (2011)
- Connolly, C.: The determination of next best views. In: *Proc. IEEE Int. Conf. on Robotics and Automation*, vol. 2, pp. 432–435. St. Louis, MO, USA (1985)
- Delmerico, J., Isler, S., Sabzevari, R., Scaramuzza, D.: A comparison of volumetric information gain metrics for active 3d object reconstruction. *Autonomous Robots* **42**(2), 197–208 (2018)
- Doumanoglou, A., Kouskouridas, R., Malassiotis, S., Kim, T.K.: Recovering 6d object pose and predicting next-best-view in the crowd. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3583–3592 (2016)
- Gschwandtner, M., Kwitt, R., Uhl, A., Pree, W.: Blesor: Blender sensor simulation toolbox. In: *International Symposium on Visual Computing*, pp. 199–208. Springer (2011)
- Hardouin, G., Morbidi, F., Moras, J., Marzat, J., Mouadib, E.M.: Surface-driven next-best-view planning for exploration of large-scale 3d environments. In: *21st IFAC World Congress (VIRTUEL)* (2020)
- Hepp, B., Dey, D., Sinha, S.N., Kapoor, A., Joshi, N., Hilliges, O.: Learn-to-score: Efficient 3d scene exploration by predicting view utility. In: *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 437–452 (2018)
- Hornung, A., Wurm, K.M., Bennewitz, M., Stachniss, C., Burgard, W.: OctoMap: An efficient probabilistic 3D mapping framework based on octrees. *Autonomous Robots* (2013). DOI 10.1007/s10514-012-9321-0. URL <http://octomap.github.com>
- Johns, E., Leutenegger, S., Davison, A.J.: Pairwise decomposition of image sequences for active multi-view recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3813–3822 (2016)
- Jovančević, I., Larnier, S., Orteu, J.J., Sentenac, T.: Automated exterior inspection of an aircraft with a pan-tilt-zoom camera mounted on a mobile robot. *Journal of Electronic Imaging* **24**(6), 061110 (2015)
- Julian, B.J., Karaman, S., Rus, D.: On mutual information-based control of range sensing robots for mapping applications. *The International Journal of Robotics Research* **33**(10), 1375–1392 (2014)
- Kavraki, L.E., Svestka, P., Latombe, J.C., Overmars, M.H.: Probabilistic roadmaps for path planning in high-dimensional configuration spaces. *IEEE Transactions on Robotics and Automation* **12**(4), 566–580 (1996)
- Khalfaoui, S., Seulin, R., Fougerolle, Y.D., Fofi, D.: An efficient method for fully automatic 3d digitization of unknown objects. *Computers in Industry* **64**(9), 1152–1160 (2013)
- Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014)
- Kriegel, S., Rink, C., Bodenmüller, T., Narr, A., Suppa, M., Hirzinger, G.: Next-best-scan planning for autonomous 3d modeling. In: *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 2850–2856. IEEE (2012)
- Kriegel, S., Rink, C., Bodenmüller, T., Suppa, M.: Efficient next-best-scan planning for autonomous 3d surface

- reconstruction of unknown objects. *Journal of Real-Time Image Processing* **10**, 611–631 (2015)
20. Lauri, M., Pajarinen, J., Peters, J., Frintrop, S.: Multi-sensor next-best-view planning as matroid-constrained submodular maximization. *IEEE Robotics and Automation Letters* **5**(4), 5323–5330 (2020)
 21. LaValle, S.M., J. J. Kuffner, J.: Randomized kinodynamic planning. *The International Journal of Robotics Research* **20**(5), 378–400 (2001). DOI 10.1177/02783640122067453
 22. Martinez-Carranza, J., Calway, A., Mayol-Cuevas, W.: Enhancing 6d visual relocalisation with depth cameras. In: *Intelligent Robots and Systems (IROS)*, 2013 IEEE/RSJ International Conference on, pp. 899–906. IEEE (2013)
 23. Mendoza, M., Vasquez-Gomez, J.I., Taud, H.: Nbv classification dataset. <https://www.kaggle.com/miguelmg/nbv-dataset> (2018). [Online; accessed 20-January-2019]
 24. Mendoza, M., Vasquez-Gomez, J.I., Taud, H.: Nbv regression dataset. <https://github.com/irvingvasquez/nbv-regression-dataset> (2018). [Online; accessed 20-January-2019]
 25. Mendoza, M., Vasquez-Gomez, J.I., Taud, H., Sucar, L.E., Reta, C.: Supervised learning of the next-best-view for 3d object reconstruction. *Pattern Recognition Letters* (2020)
 26. Monica, R., Aleotti, J.: Contour-based next-best view planning from point cloud segmentation of unknown objects. *Autonomous Robots* **42**(2), 443–458 (2018)
 27. Moritani, R., Kanai, S., Date, H., Niina, Y., Honma, R.: Plausible reconstruction of an approximated mesh model for next-best view planning of sfm-mvs. *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences* **43**, 465–471 (2020)
 28. Mur-Artal, R., Montiel, J.M.M., Tardos, J.D.: ORB-SLAM: a versatile and accurate monocular SLAM system. *IEEE Transactions on Robotics* **31**(5), 1147–1163 (2015). DOI 10.1109/TRO.2015.2463671
 29. Potthast, C., Sukhatme, G.: A probabilistic framework for next best view estimation in a cluttered environment. *J. Vis. Comun. Image Represent* **25**(1), 148–164 (2014)
 30. Ramanagopal, M.S., Nguyen, A.P.V., Ny, J.L.: A motion planning strategy for the active vision-based mapping of ground-level structures. *IEEE Transactions on Automation Science and Engineering* **15**(1), 356–368 (2018)
 31. Scott, W., Roth, G., Rivest, J.: View planning for automated three-dimensional object reconstruction and inspection. *ACM Computing Surveys* **35**, 64–96 (2003). DOI 10.1145/641865.641868
 32. Song, S., Jo, S.: Online inspection path planning for autonomous 3d modeling using a micro-aerial vehicle. In: *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6217–6224. IEEE (2017)
 33. Song, S., Jo, S.: Surface-based exploration for autonomous 3d modeling. In: *IEEE International Conference on Robotics and Automation*, pp. 4319–4326. IEEE (2018)
 34. Themistocleous, K., Ioannides, M., Agapiou, A., Hadjimitsis, D.G.: The methodology of documenting cultural heritage sites using photogrammetry, uav, and 3d printing techniques: the case study of asinou church in cyprus. In: *Third International Conference on Remote Sensing and Geoinformation of the Environment (RSCy2015)*, vol. 9535. International Society for Optics and Photonics (2015)
 35. Thrun, S., Burgard, W., Fox, D.: *Probabilistic Robotics*. The MIT Press (2005)
 36. Torabi, L., Gupta, K.: An autonomous six-dof eye-in-hand system for in situ 3d object modeling. *International Journal of Robotics Research* **31**(1), 82–100 (2012)
 37. Vasquez-Gomez, J.I., Sucar, L.E., Murrieta-Cid, R.: View/state planning for three-dimensional object reconstruction under uncertainty. *Autonomous Robots* **41**(1), 89–109 (2017)
 38. Wang, Y., James, S., Stathopoulou, E.K., Beltrán-González, C., Konishi, Y., Del Bue, A.: Autonomous 3-d reconstruction, mapping, and exploration of indoor environments with a robotic arm. *IEEE Robotics and Automation Letters* **4**(4), 3340–3347 (2019). DOI 10.1109/LRA.2019.2926676
 39. Wu, C., Zeng, R., Pan, J., Wang, C.C., Liu, Y.J.: Plant phenotyping by deep-learning-based planner for multi-robots. *IEEE Robotics and Automation Letters* **4**(4), 3113–3120 (2019)
 40. Wu, Z., Song, S., Khosla, A., Yu, F., Zhang, L., Tang, X., Xiao, J.: 3d shapenets: A deep representation for volumetric shapes. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1912–1920 (2015)
 41. Yuan, W., Khot, T., Held, D., Mertz, C., Hebert, M.: Pcn: Point completion network. In: *2018 International Conference on 3D Vision (3DV)*, pp. 728–737. IEEE (2018)
 42. Zeng, R., Wen, Y., Zhao, W., Liu, Y.J.: View planning in robot active vision: A survey of systems, algorithms, and applications. *Computational Visual Media* **6**(3), 225–245 (2020)
 43. Zeng, R., Zhao, W., Liu, Y.J.: Pc-nbv: A point cloud based deep network for efficient next best view planning. In: *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (2020)