

BSR-NBV: Breaking Stagnation with Random Pose Refinement in Probabilistic Grids for Latent Space Next-Best-View Prediction

Saulo A. Gante-Díaz , Juan Irving Vasquez , and Erik Zamora 

Abstract—The Next-Best-View (NBV) problem is fundamental to autonomous 3D object reconstruction, where the goal is to generate a complete 3D model by sequentially sensing a target's surface. This paper proposes a novel latent space-based method to address this challenge. Our architecture is trained in two distinct stages: the first stage learns to encode a probabilistic grid into a compact latent representation, while the second stage performs continuous NBV regression. Furthermore, we incorporate a pose refinement mechanism to mitigate stagnation—a phenomenon where the network repeatedly predicts views with low information gain that fail to update the probabilistic grid. Experimental evaluations on the Google Scanned Objects dataset demonstrate that the proposed method achieves superior surface coverage compared to state-of-the-art approaches.

Link to graphical and video abstracts, and to code:
<https://latam.ieeer9.org/index.php/transactions/article/view/10538>

Index Terms—Next-best-view, view planning, 3D object reconstruction, probabilistic grid, 6d pose.

I. INTRODUCTION

THE Next-Best-View problem (NBV) represents a challenge in computer vision, more specifically in the field of 3D reconstruction. This problem arises from the limited sensor's field of view, so it has to be moved to different poses to capture the complete surface of the object. The primary objectives are to determine the best sensor pose for acquiring missing information and to achieve this using a minimal set of views. Rather than relying on data capture from fixed predefined positions, NBV algorithms prioritise selecting views that reduce uncertainty, increase surface coverage, and minimise reconstruction errors.

The concept of NBV was first introduced in the mid-1980s by Connolly [1], who proposed a couple of algorithms that allowed a sensor to move freely in a view space by processing an octree to select the NBV. On its part, in [2] an occlusion avoidance system was proposed to select views on a spherical surface which is assumed to be around the target. Showing the first steps to solve this task, this concept has been widely

explored for different tasks like object reconstruction [3], [4], robot grasping [5]–[7], or environment exploration [8].

State-of-the-art methods typically rely on processing of partial point clouds (onwards PC) [9] or probabilistic grids [10] to compute the NBV. These strategies usually formulate the problem as a classification task over a predefined set of candidate poses. An important limitation of these approaches is the implicit assumption that the set of candidates (and consequently the NBV) are sufficient to ensure complete surface coverage. Consequently, they often exhibit a limited generalisation when confronted with the diverse geometric variations of the objects.

In this paper, we present a novel NBV method that employs a compact latent representation [11] of the probabilistic grid in order to generalize the performance across object shapes. Our approach combines this latent representation with a regression stage for free space pose prediction, alongside a pose refinement stage to mitigate stagnation issues associated with the limited number of training examples. When we refer to stagnation, we mean that the neural network gets stuck predicting the same pose in each iteration, as the probabilistic grid is not improved by the predicted pose. The key contributions are summarized as follows:

- The proposition of a latent representations of the object to predict the NBV.
- A novel neural network architecture that combines an autoencoder for code generation combined with a multilayer perceptron to predict the NBV in the continuous space.
- The proposition of a refinement mechanism that translates the viewpoint into a uniform distribution of views given a neighbourhood near the predicted viewpoint.

We experimentally evaluate the efficacy of the method on well known dataset [12] in terms of surface coverage. The results indicate that planning within the latent space yields a performance competitive with approaches that use explicit representations. Furthermore, the proposed method outperforms previous state-of-the-art approaches on several object geometries.

Our paper is organised into six sections. Section I introduces the problem and outlines the research objectives. Section II provides a review of related literature. Section III states the problem and its boundaries. Section IV describes the proposed methodology. Section V presents the experiments performed, evaluations to show the capabilities of the method, and discusses the results. Finally, section VI concludes the paper and suggests lines for future work.

The associate editor coordinating the review of this manuscript and approving it for publication was Eduardo José da Luz (*Corresponding author: Saulo Abraham Gante*).

J. I. Vasquez is with Instituto Politécnico Nacional, Centro de Innovación y Desarrollo Tecnológico en Cómputo, Ciudad de México, México (e-mail: ivasquez@ccc.inaoep.mx).

Saulo Abraham Gante, and E. Zamora are with Instituto Politécnico Nacional, CIC. Av. Juan de Dios Batiz S/N, Col. Nueva Industrial Vallejo, Gustavo A. Madero, 07738, Ciudad de México, México (e-mails: sganted2023@cic.ipn.mx, and ezamorag@ipn.mx).

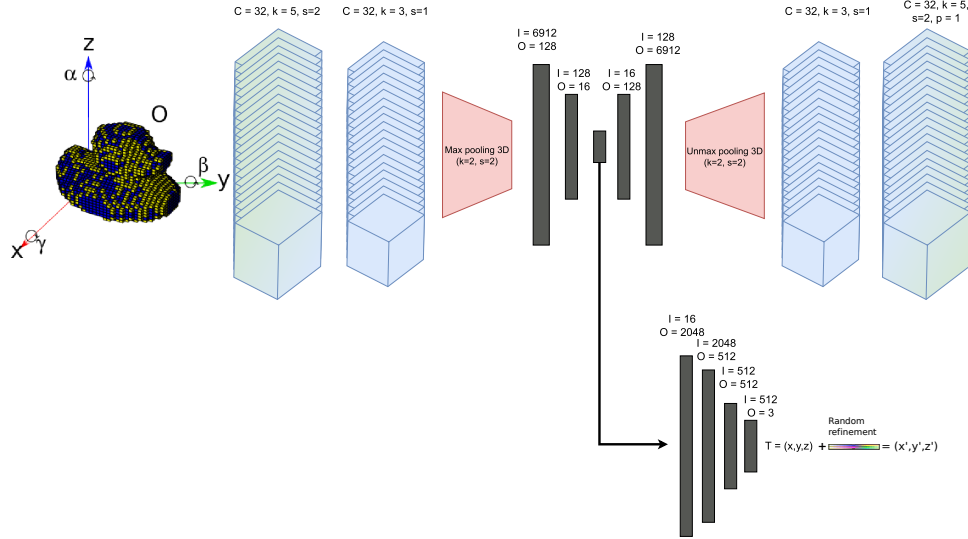


Fig. 1. Neural network architecture overview. The duck probabilistic grid in the left is the partial model at current reconstruction state. Our proposed neural network BSR-NBV makes use of an autoencoder for self-supervise the probabilistic grid dimensionality reduction. The reduced representation, latent space, is used as input in a multilayer perceptron to predict the NBV and a random restricted pose displacement is done to avoid stagnation issues.

II. RELATED WORK

A. Next-Best-View Estimation

The NBV problem, a special instance of the view planning problem, has been widely explored with a variety of proposed methodologies to solve it. According to [13], although the problem has been widely studied, there are still many unsolved challenges. When talking about object reconstruction NBV, most of the issues are related to the object variables, such as textures, auto-occlusions, surface materials, and environmental lighting, which increase the difficulty of generalisation.

In [14], they propose a NBV prediction strategy for PC's, where features are extracted from the object's PC, processed, and then fed to an attention unit, and the output is processed by an MLP to select it among N positions.

Following a similar approach for model processing, the work in [10] proposes a probabilistic grid representation to capture the geometry of the reconstructed volume; which is processed by a neural network with 3D convolutional layers and at the output linear layers for prediction in a discretised space in 14 visualisation poses in the semi-sphere, and each pose visualises the center of the object.

Then in [15], the authors improve the previous work by avoiding space discretization. Their model operates in a continuous space, predicting the three-dimensional spatial coordinates of the sensor, (x, y, z) , and with these positions they can predict the orientations in yaw and pitch.

Some recent proposals opt for the use of one-shot strategies, where a model is generated from a number of initial visualisations and is processed to plan all the views necessary to complete a reconstruction. This is the case of [16], where long-tailed sampling is used. It starts with a random view and a NBV algorithm adds one to four views. Then, a model is created with this information, and finally, it is processed by a neural network that outputs a 32-bit vector corresponding to the possible poses to be selected. In [17] the use of this

methodology is applied for object recognition and classification, where the analysed data is an image of an object; it is expected to perform an estimation of the NBV given the appearance of the object and its surface orientation.

A topological space analysis is presented in [18], where the authors propose an information gain metric based on the PC topology with which they reconstruct the space to train a reinforcement learning model; the value assigned to each view will be based on the topological attributes obtained from a distance, and the reward is calculated with the value of the view, considering the PC at a current time and the previous one. Using a multilayer perceptron (MLP), the method predicts 200 possible poses. Furthermore, they do not report completeness metrics for the reconstructions. In [9] a NBV approach based on PC's is proposed, the system uses a partial PC of 1024 randomly sampled points and a view state vector that encodes viewpoints already visited. Then it provides the NBV among 33 viewpoints.

From this overview, it is clear that most of the previous works address the problem is viewed from a classification perspective, restricting the solution to a set of viewpoints. This approach assumes that this pose will provide sufficient information to reconstruct the object, failing to consider whether self-occlusions might prevent the observation of the object. We consider that a method for predicting the NBV in free space is required to relax viewpoint restrictions in order to reconstruct objects with complex shapes.

B. Deep Learning-based Pose Estimation

Gao et al. [19] mention that some works propose the discretization of a continuous space of poses, allowing them to formulate the problem as a classification [20], [21]; some others look for prediction proposals in a continuous space [22]–[26] and propose the use of quaternion representations

to predict rotations [27]–[29] or rely on the axial-angular representation.

For its part, in [30] a comparison between different ways of representing rotations in a 3D space is provided, suggesting that neural networks are better at approximating continuous rotation spaces. Since Euler and quaternion representations have discontinuities, they can cause problems during the learning process.

III. NBV PREDICTION IN LATENT SPACE

In this section, we will provide a description of the proposed method. First, we provide a formal statement of the problem, and then an overview of the NBV method. Then, the proposed network architecture and pose refinement are explained.

A. Problem Statement

The NBV problem is formulated as finding the best position for a sensor using accumulated information that is stored in a data structure (3D model) which is usually composed of PC's or probabilistic grids. This process is repeated iteratively until a requirement is satisfied, information gain and number of visited views, for example. The resulting sequence of views $\{v_1, \dots, v_n\} \subset V$ belongs to the view space $\mathcal{V} = \mathbb{R}^3 \times SO(3)$, where $SO(3) = \{R \in \mathbb{R}^{3 \times 3} | RR^T = I_{3 \times 3}, \det R = 1\}$ is the special orthogonal group in $\mathbb{R}^{3 \times 3}$.

When working with probabilistic grids, the set of views is evaluated by the information gained at each one, and it can be expressed in equation 1.

$$G_v = \sum_{r \in R_v} \sum_{x \in X} I \quad (1)$$

where R_v represents the set of rays in the viewing pose, while X represents the set of voxels in the trajectory of the viewing ray, the information is defined by the Shannon entropy of the volumetric information contained in a voxel that is visible in a particular view [13], namely,

$$I = -P_o(x) \ln P_o(x) - \bar{P}_o(x) \ln \bar{P}_o(x) \quad (2)$$

Where $P_o(x)$ is the probability of the existence of an object in the voxel, while $\bar{P}_o(x) = 1 - P_o(x)$ is the probability that no object exists in the voxel. The set of views that cover the object is defined by the views that maximise the information gain, formally described in equation 3.

$$V^* = \arg \max_{v \in V} \mathbb{E}[G_v | V] \quad (3)$$

All selected views are integrated into a single model, resulting in an object reconstruction that is expected to fully cover the target object.

B. NBV Regression in Latent Space

For view computation, neural network-based methods can be grouped into classification and regression categories. In the former, a sphere of radius r is defined, on which a set of discrete points are set to select NBV. In the latter, the

network predicts a full continuous pose. This work adopts the regression paradigm, defining the predictor as follows:

$$f_{NBV}(M) : \mathbb{R}^{h \times w \times d} \rightarrow \mathbb{R}^3 \times SO(3), \quad (4)$$

So, as in equation 4 the problem lies on finding a function, f_{NBV} , which by processing a model, M , of dimensions (h, w, d) could predict the position $\mathbb{R}^3$ and the orientation represented with a rotation matrix in the special orthogonal group three, $SO(3)$. Note that M is a partially reconstructed model.

We propose the use of M to learn how to reconstruct the object, see equation 5. Therefore M is introduced into an encoder, $m = f(M)$, and the output code goes through a decoder, $g(m)$, and produces a reconstruction of the input, $\hat{M}$, namely:

$$\hat{M} = g(f(M)), \quad (5)$$

where

$$\hat{M} - M \approx 0 \quad (6)$$

After training, the code, m , is used as input to an MLP,

$$\hat{y}_{NBV} = h(m). \quad (7)$$

For the autoencoder training, the target is to minimize the MSE:

$$\mathcal{L}_M(M, \hat{M}) = \frac{1}{n} \sum_{i=1}^n (M_i - \hat{M}_i)^2, \quad (8)$$

and for the NBV regression, the target is to minimize:

$$\mathcal{L}_{NBV}(y, \hat{y}) = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (9)$$

C. Neural Network Architecture

To explain the proposed neural network architecture, the following notation is provided: $C3(C, k, s, p)$ is for 3D convolution, $TC3(C, k, s, p)$ is for 3D transposed convolution, where C is the number of channels or filters, k is the kernel size, s is stride, p is the padding, $FC(n)$ is a fully-connected layer with n neurons, $MP(k, s)$ is the max-pooling operation, $UMP(k, s)$ is the unmax-pooling operation, both in 3D.

1) *Reconstruction Architecture*: The proposed architecture consists of an autoencoder that is fed with 3D probabilistic grids of dimension $31 \times 31 \times 31$, in the following step, the code is feed to an MLP to predict the NBV. The autoencoder consists of two parts; the encoder has two 3D convolution layers, a max-pooling, and two fully-connected layers; the decoder consists of two fully-connected layers, a 3D unmax-pooling, and two 3D transposed convolution layers. On the other hand, the MLP consists of four fully-connected layers. See Fig. 1. To summarize, the architecture for the autoencoder is as follows: $C3(32, 5, 2, 0)$, $C3(32, 5, 2, 0)$, $MP(2, 2)$, $FC(6912)$, $FC(128)$, $FC(n)$, $FC(128)$, $FC(6912)$, $UMP(2, 2)$, $TC3(32, 3, 1, 0)$, $TC3(32, 5, 2, 1)$ where n is the latent space dimension that can be adjusted and is analysed during experimentation.

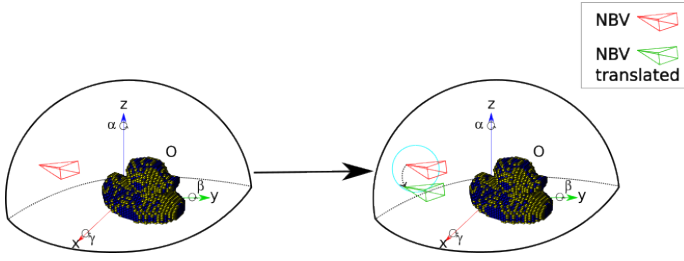


Fig. 2. NBV refinement. The view is translated into one neighbour view, given a spherical cap neighbourhood (in light blue).

2) *NBV Regression Architecture*: Looking to find the function f_{NBV} , we believe that it can be easily approximated by an MLP, as it is well-known that it serves as a good function approximator. For the MLP the input is of size n and the architecture goes as follows: $FC(2048)$, $FC(512)$, $FC(512)$, $FC(3)$; here we were inspired by the architecture presented in [30]. The activation functions used are ReLU for the FC layers in the autoencoder, LeakyReLU for the C3, TC3, and FC layers of the MLP and hyperbolic tangent for the output layer of the MLP.

D. Pose Refinement

Despite the network predicting the NBV in a 3D space, when working with probabilistic grids and neural networks, problems arise when there is no substantial information gained to add to the model; the view overlap is mostly part of the known space. This problem causes neural networks to converge; as a result, the NBV is repeated indiscriminately.

To address the limitations of neural networks, we propose a pose refinement stage, which refines the network output $v_{pred} = (x, y, z)$. Considering that our view space is restricted to the surface of the superior hemisphere of a sphere, $S^2(R) = \{x \in \mathbb{R}^3 : \|x\| = R\}$, for each predicted NBV, v_{pred} , a spherical cap with radius r is considered to transform it to v'_{pred} , as shown in Fig. 2. The v_{pred} is considered to be the centre of the spherical cap.

Using polar coordinates, we uniformly sample points in a uniform distribution $\mathcal{U}$, $\phi_i = \mathcal{U}(0, 2\pi)$, while for θ a $u_i = \mathcal{U}(0, 1)$ is performed to avoid a concentration at the pole. Obtaining θ_i with:

$$\theta_i = \cos^{-1}(\cos(r) + u_i(1 - \cos(r))), \quad (10)$$

where, the point in polar coordinates is defined as $p_{polar}(\phi, \theta)$. In order to transform its coordinates to Cartesian coordinates, we have a point p_{cart} expressed as:

$$p_{cart}(x, y, z) = \begin{bmatrix} \sin \theta_i \cos \phi_i \\ \sin \theta_i \sin \phi_i \\ \cos \theta_i \end{bmatrix} \quad (11)$$

Then, with the Rodrigues rotation [31], we have our point rotated as follows:

$$p' = R\left(\frac{v_{pred}}{\|v_{pred}\|}\right)p, \quad (12)$$

ensuring that the point is contained in the surface, we scale it to the radius R . Consequently, for a point $p \in \mathbb{R}^3$, and an angle limit $r \in (0, \pi]$, we define our spherical cap as follows:

$$S_c^2 = \mathcal{U}(\{x \in \mathbb{R}^3 : \|x\| = R, \arccos\left(\frac{p}{\|p\|} \cdot \frac{x}{R}\right) \leq r\}) \quad (13)$$

As illustrated in Fig. 2, this pose refinement ensures that the translated pose is in the view space and in a neighbourhood near the predicted NBV. By performing this translation, we expect to increase the information gained per view; making it useful for cases where the lacks of information causes stagnation.

IV. EXPERIMENTS

In this section, we present experiments to show the effectiveness of the proposed code-based planning and refinement approach. First, we describe the dataset, then we present the results obtained during training to examine the performance of the proposed network architectures. Experimentation was conducted using a NVIDIA Tesla P100 GPU provided by Google Colab.

Then, a test experiment is conducted to evaluate the performance of the proposed method versus similar NBV approaches on previously unseen objects. An ablation study is performed to evaluate the performance of network NBV prediction. These experiments were executed into a Ryzen 7 with an NVIDIA GeForce RTX 4090 GPU. At the end of the section, we present a discussion of the results and limitations.

A. Dataset Description

The dataset we are using was proposed in [10]. The dataset consists of 15,166 records and is publicly available, already partitioned by the authors into 80-20 training and validation sets. Each consisting of objects in different stages of reconstruction that we call models, M , those models are partially reconstructed, and their corresponding NBV, $v = (x, y, z, \alpha, \beta, \gamma)$, which is a vector that expresses its position and orientation. The models are codified in a probabilistic grid and are of size $32 \times 32 \times 32$. Here we re-scaled the models to $31 \times 31 \times 31$ size, to fit into the autoencoder architecture; to achieve the re-scaling, trilinear interpolation was performed.

B. Training

Our training strategy consists of two steps; at first, we optimise the autoencoder parameters with the main objective of learning a code meaningful for reconstruction and for the NBV prediction.

1) *Autoencoder Training*: We perform a set of experiments on the proposed architecture with minimal variations. We experiment with different code sizes, z , using the whole probabilistic dataset. The code sizes are as follows: 64, 32, 16, 8 and 4.

To minimise the error described in Equation 8, we use the Adam optimiser, which automatically adjusts the learning rate. Training was carried out over 500 epochs with learning rate

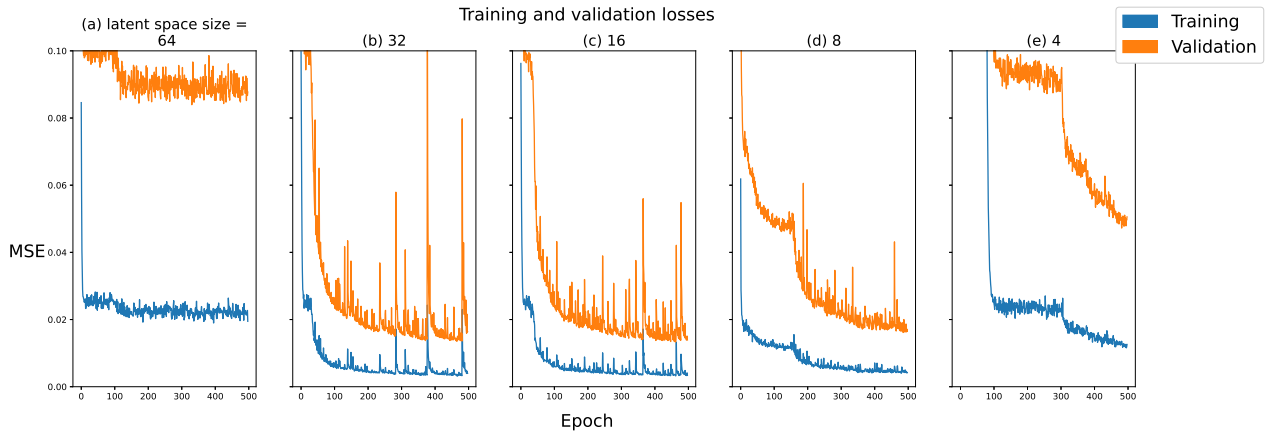


Fig. 3. Training and validation losses of autoencoder training. The autoencoder was trained with different latent spaces (ls) in order to explore how much it was possible to reduce the dimensionality of the representation and obtain a reconstruction that resemble to the input. training and validation losses for a ls dimension of a) 64 elements, b) 32 elements, c) 16 elements, d) 8 elements, and e) 4 elements. These plots suggest that (a), (d), and (e) does not generate a useful latent representation for the reconstruction process.

TABLE I
VALIDATION MINIMUM LOSS OBTAINED PER MODEL
DURING TRAINING

Treatment	Latent Space Size	MSE Loss
Autoencoder	64	0.01860
	32	0.01146
	16	0.01007
	8	0.01080
	4	0.01675

TABLE II
MLP SEARCH THROUGH DIFFERENT LEARNING RATE
VALUES

Treatment	Learning Rate	MAE
MLP	1×10^{-3}	1.488
	1×10^{-4}	1.513
	1×10^{-5}	1.449

5×10^{-3} and a batch size equal to 1000. The whole network training took an average time of seven hours.

We can see from Fig. 3 that the larger and smaller sizes of the code do not perform well compared against the other sizes. We summarise the results considering only the best performance in table I. In addition, it is possible to see that the use of z size of 16 has a better performance in terms of lower MSE. See Fig 3(c); also, it shows that b and d have good performances.

2) *MLP Training*: In a similar fashion, we look to minimise the equation 9 and we use the Adam optimiser, perform the training over 1500 epochs, and a batch size of 500. For this experiment, we modified the learning rate in 1×10^{-N} for $N = 3, 4, 5$. The authors in [15] suggest adding an MAE metric to measure how well predictions in Euclidean space are, so we add it in validations during the training process. Fig. 4 shows the performance of the proposed architecture during training and validation.

Summarizing this experiment, the models showed similar performance during training; since we want to choose the

model that performs better in Euclidean space, we took the results over this space to make a selection, where the use of a learning rate of 1×10^{-5} in the MAE metric got the best results; see Table II.

C. Test

In this section, we provide our test results. We took a subset of 227 objects from the Google scanned objects dataset (GSOD) [12].

We used a simulated 3D environment for the view planning process; an unknown object is put into the simulated scene and we expect that the methods can predict the NBV. We expect the object to be centred in the frame of reference at $(\pm x, \pm y, \pm z)$, and contained in a bounding box with sides equal to 0.4. We use Open3D [32] to simulate the world and sensors. The input to our method system is a probabilistic occupancy grid dynamically updated from the simulation at each step, whose latent representation is fed into the MLP to predict the next-best viewpoint. During testing, there is no predefined target pose; performance is assessed solely based on surface coverage to quantify the completeness of the 3D reconstruction.

To compare the method's performance, we consider having a partial model, P_M , and a ground truth PC, P_{GT} , defining the surface coverage to quantify the ground truth PC points that are adequately represented by the partial PC, see equation 14.

$$Cover(P) = \frac{1}{|P_{GT}|} \sum_{p_m \in P_M} U(\min_{p \in P_{GT}} \|p_m - p\|_2 - \epsilon), \quad (14)$$

where U is a Heaviside function and ϵ is a distance threshold.

When talking about surface coverage as a comparison reference, usually ten views are reported in similar works; here we extended the experiments since we wanted to explore if the methods that use probabilistic grids show stagnation issues by themselves. We tested our proposal on the entire set of objects. To ensure similar conditions, we initialized the reconstruction

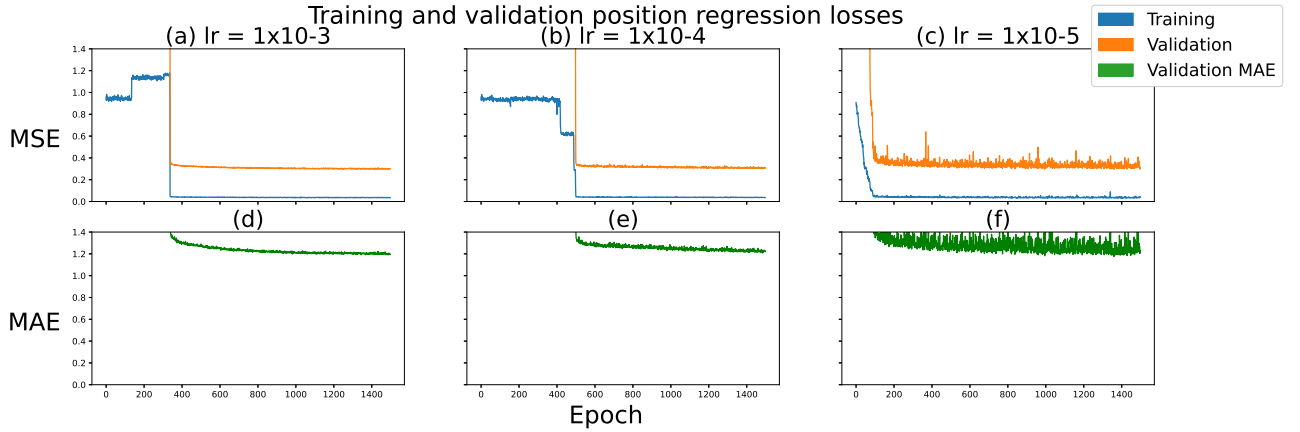


Fig. 4. Training and validation losses of MLP training for NBV regression. a),d) presents the learning rate (lr) 1×10^{-3} losses, b),e) 1×10^{-4} , and c),f) 1×10^{-5} . Training results show that the small lr (c) is useful since the MLP reduces MSE rapidly. On the other hand, validation results (d,e,f) shows that neural networks struggles with pose prediction.

TABLE III
RECONSTRUCTION MEAN AVERAGE SURFACE COVERAGE

Method	Mean Average
Our + 45°	89.40
Our + 22.5°	86.29
Our + 15°	85.34
Our + 11.5°	84.47
Our + 9°	85.08

process at the same pose for every tested method. Here we explore different angular offsets in order to validate at which angle the model tends to avoid the undesired stagnation. We experimented with 45, 22.5, 15, 11.5, and 9 degrees, see Fig. 5.

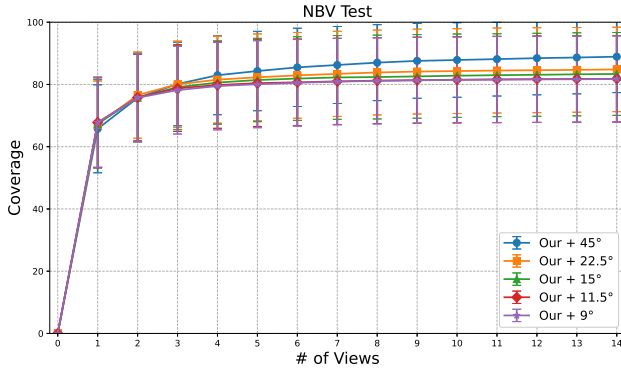


Fig. 5. Test results with pose refinement. The plot show the mean and standard deviation coverage in percentage at each step per method. The results show that our methods achieves a mean coverage of 80% at each experiment.

D. Ablation Study

Finally, we conducted an ablation study. Here, we remove the pose refinement stage to validate the impact it has on the proposed method while comparing its performance with other methods.

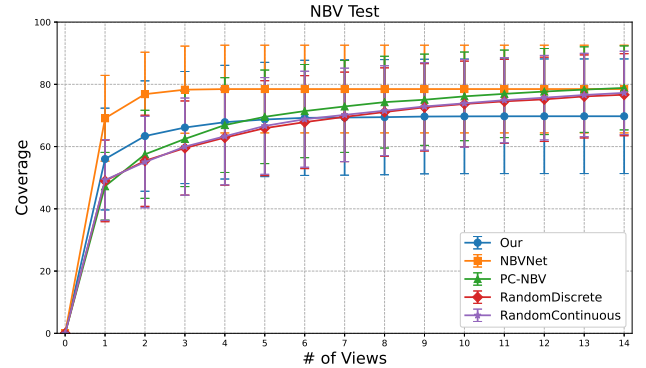


Fig. 6. Ablation results and comparison SOTA methods. The results shows that removing the pose refinement stage reduces the method mean coverage substantially.

From Fig. 6, we observe that the proposed network, without pose refinement, does not outperform the baselines; see Table III. In particular, we find that relying on probabilistic grids for NBV tasks causes the method to suffer from stagnation issues when an information gain threshold is not met; as shown in Fig. 6 with the coverage curves of our proposal.

E. Reconstruction Performance on Additional Datasets

To demonstrate that BSR-NBV, trained on [10], generalizes to different and complex scenarios, we conducted experiments on two datasets: the commonly used objects in NBV benchmarks, and a subset from the GSOD dataset. For the NBV benchmark objects, each object was reconstructed 10 times to confirm that our method does not rely solely on randomness, but rather leverages it as a complementary mechanism within the pose refinement stage. For the GSOD subset, experiments were conducted once per object due to computational constraints; therefore, statistical testing is not available for this group and relative improvement ($\Delta\%$) is reported instead.

The object names and mean AUC surface coverage results are presented in Table V, where we also report the relative improvement between the best of our methods and the state-

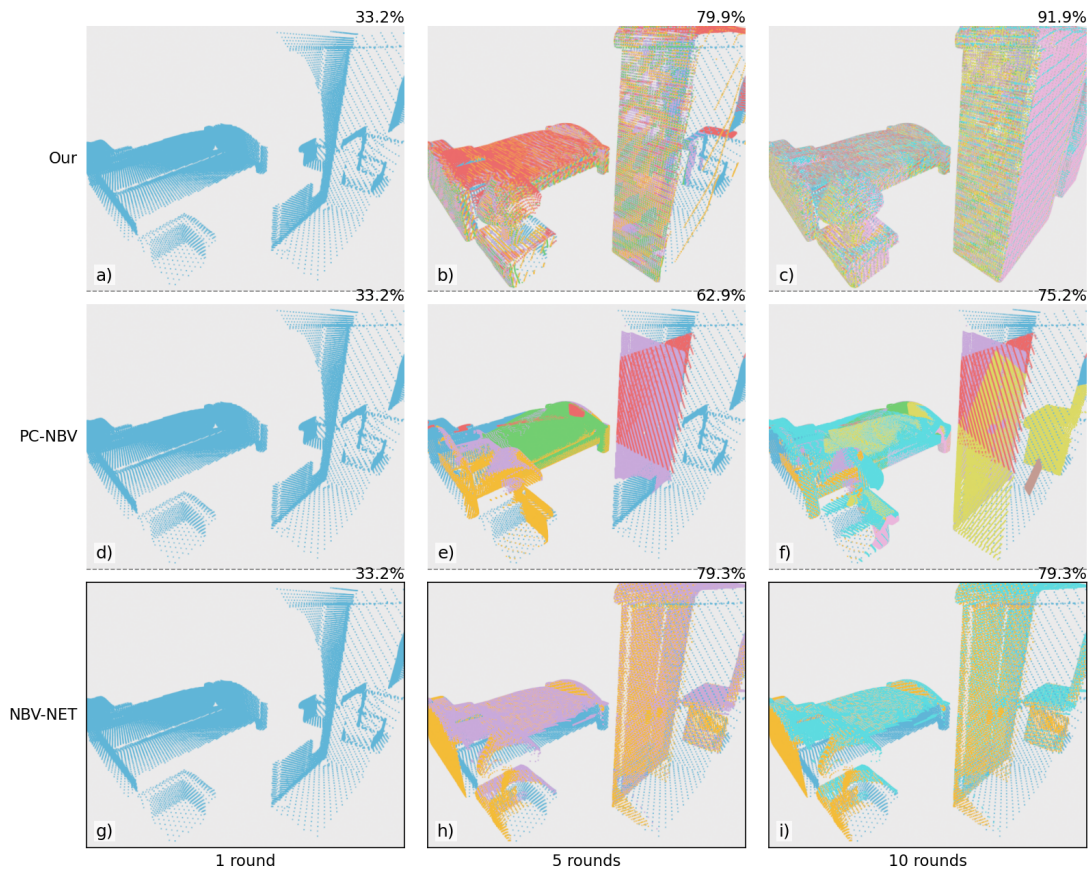


Fig. 7. Reconstruction progress comparison on the "Children Bedroom" model. The surface coverage is shown in the upper right corner. BSR-NBV selects better NBV's on this geometrically complex model. a),d),g) Presents the initial partial models for each method, all processes start with the same view, b),e),h) Partial reconstructed objects at the 5th round, our method (b) cover the 79.9% of the object by this round, c) Complete reconstructed objects at 10th rounds, our method performs a coverage of the 91.9% of the object while i) NBV-NET suffers from stagnation as the coverage is the same as in the 5th round, f) PC-NBV increases the coverage until 75.2% struggling to completely reconstruct the object.

of-the-art baselines, along with Welch's t-test p-values where available. Table VI reports the mean Chamfer distance at view 10, which measures how closely the reconstructed surface approximates the ground truth. Results show that BSR-NBV outperforms the baseline methods in the majority of the evaluated objects, achieving higher AUC surface coverage and lower Chamfer distance, with relative improvements of up to 31.9% over the best competing method. The results across both datasets confirm that the stochastic pose refinement stage effectively complements the learned policy, as the method maintains a competitive advantage even on complex and unseen objects.

Fig. 7 shows a qualitative reconstruction comparison on the "Children Bedroom" object, illustrating how BSR-NBV acquires more surface information in fewer views compared to the baselines.

F. Discussion

We took into consideration the SOTA NBV methods proposed in [9], [10] for PC's and octomaps, respectively. In addition, as reported in the literature, we use two random NBV selectors, a continuous and a discrete one. All methods were

evaluated using the authors' original implementations and pre-trained weights. To ensure a fair comparison and eliminate potential sources of bias, we enforced consistent constraints throughout testing, such as initiating the reconstruction from identical poses.

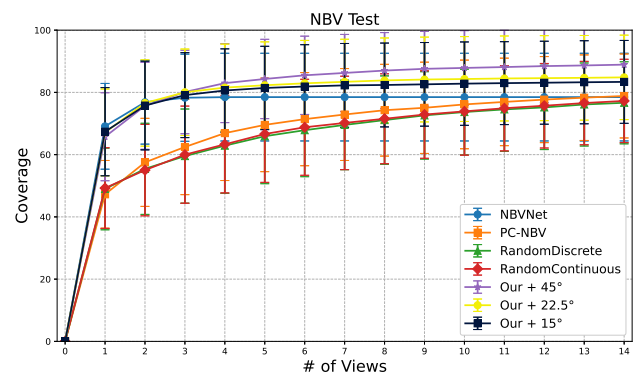


Fig. 8. Results and comparison with reported methods. The comparison shows that our method improves the mean coverage considering SOTA and random methods. NBV-Net uses probabilistic grids and shows the stagnation issues as reported in our findings.

TABLE IV
RECONSTRUCTION MEAN AVERAGE SURFACE COVERAGE
AND # OF OBJECTS WITH THE BEST PERFORMANCE IN
COVERAGE

Method	Mean Average	# of objects
NBVNet [10]	78.48	28
PC-NBV [9]	76.13	14
Random Discrete	73.68	3
Random Continuous	73.89	2
Our + 45°	89.40	99
Our + 22.5°	86.29	35
Our + 15°	85.34	22
Our + 11.5°	84.47	13
Our + 9°	85.08	11

The mean coverage results, presented in Fig. 8, demonstrate that the pose-refinement variants perform competitively in the GSOD. This performance suggests that the refinement process effectively enhances the quality of the reconstruction compared to baseline approaches; in most cases, showing superior results.

The addition of refinement helps the model to perform better, suggesting that the model tends to predict NBV poses that do not gain a considerable amount of information, causing it to converge; on average, it makes the model gain 16.47% more information, and in the best case 19.76%, suggesting that these modifications effectively address the stagnation issues observed in the baseline method, as seen in the ablation study.

To summarize results, we provide Table IV, where the average mean surface coverage at the tenth view and the number of objects where the method performs the best are reported in the GSOD.

In addition, Table IV reports the number of objects in which the coverage performance is the best of all compared methods. The results showed that our method in the best case performed better on 99 objects.

V. CONCLUSIONS

This study proposed code-based (latent space) guidance with pose refinement for NBV computation, a novel supervised learning method for planning free space NBV points designed for 3D reconstruction using probabilistic grids. The main components of the method are a model compaction stage driven by an autoencoder architecture, a set of fully connected layers to perform NBV regression, and finally a pose refinement stage to avoid stagnation. As a result, the complete method showed superior performance in reconstruction coverage compared to conventional deep learning SOTA methods.

We found that current neural networks suffer from stagnation issues when working with probabilistic grids due to the nature of the representation when insufficient new information is captured from the environment, and the grid is not updated, which causes the networks to predict the same NBV perspective in an endless loop. This is probably due to poor generalisation caused by the insufficient number of examples in the dataset.

The proposed method emerges as a novel and promising approach for NBV selection, proving its capabilities in covering more surface area on complex objects compared to

existing state-of-the-art methods. Despite the merits of the method, several challenges remain in improving these types of approach. The addition of a pose refinement stage is useful in preventing to the neural network from stagnating during the reconstruction process.

We believe that future work needs to expand research on neural network architectures to improve performance while operating probabilistic grids, create larger and diverse NBV probabilistic grid datasets, 6-DOF NBV prediction, and train the architecture at the same time.

ACKNOWLEDGMENTS

This work was supported by the Centro de Investigación en Computación - Instituto Politécnico Nacional (SIP-20240108, SIP-20253734, SIP-IND-2026-0031, SIP-20240975, SIP-20250200), and the Mexican National Secretary of Science, Humanities, Technology and Innovation SECIHTI under the doctoral grant 2020-2025.

REFERENCES

- [1] C. Connolly, "The determination of next best views," in *Proceedings. 1985 IEEE international conference on robotics and automation*, vol. 2. IEEE, 1985, pp. 432–435, doi: 10.1109/ROBOT.1985.1087372.
- [2] S. Sakane, M. Ish, and M. Kakikura, "Occlusion avoidance of visual sensors based on a hand-eye action simulator system: Heaven," *Advanced robotics*, vol. 2, no. 2, pp. 149–165, 1987, doi: 10.1163/156855387X00138.
- [3] T. Wang, W. Xi, Y. Cheng, H. Han, and Y. Yang, "RI-nbv: A deep reinforcement learning based next-best-view method for unknown object reconstruction," *Pattern Recognition Letters*, vol. 184, pp. 1–6, 2024, doi: 10.1016/j.patrec.2024.05.014.
- [4] Y. Han, I. H. Zhan, W. Zhao, and Y.-J. Liu, "A double branch next-best-view network and novel robot system for active object reconstruction," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 7306–7312, doi: 10.1109/ICRA46639.2022.9811769.
- [5] M. Breyer, L. Ott, R. Siegwart, and J. J. Chung, "Closed-loop next-best-view planning for target-driven grasping," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 1411–1416, doi: 10.1109/IROS47612.2022.9981472.
- [6] M. Naazare, F. G. Rosas, and D. Schulz, "Online next-best-view planner for 3d-exploration and inspection with a mobile manipulator robot," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 3779–3786, 2022, doi: 10.1109/LRA.2022.3146558.
- [7] D. Morrison, P. Corke, and J. Leitner, "Multi-view picking: Next-best-view reaching for improved grasping in clutter," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 8762–8768, doi: 10.1109/ICRA.2019.8793805.
- [8] R. Almadhoun, A. Abduldayem, T. Taha, L. Seneviratne, and Y. Zweiri, "Guided next best view for 3d reconstruction of large complex structures," *Remote Sensing*, vol. 11, no. 20, 2019, doi: 10.3390/rs11202440.
- [9] R. Zeng, W. Zhao, and Y.-J. Liu, "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)*. IEEE, 2020, pp. 7050–7057, doi: 10.1109/IROS45743.2020.9340916.
- [10] M. Mendoza, J. I. Vasquez-Gomez, H. Taud, L. E. Sucar, and C. Reta, "Supervised learning of the next-best-view for 3d object reconstruction," *Pattern Recognition Letters*, vol. 133, pp. 224–231, 2020, doi: 10.1016/j.patrec.2020.02.024.
- [11] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *science*, vol. 313, no. 5786, pp. 504–507, 2006, doi: 10.1126/science.1127647.
- [12] L. Downs, A. Francis, N. Koenig, B. Kinman, R. Hickman, K. Reymann, T. B. McHugh, and V. Vanhoucke, "Google scanned objects: A high-quality dataset of 3d scanned household items," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 2553–2560, doi: 10.1109/ICRA46639.2022.9811809.
- [13] R. Zeng, Y. Wen, W. Zhao, and Y.-J. Liu, "View planning in robot active vision: A survey of systems, algorithms, and applications," *Computational Visual Media*, vol. 6, pp. 225–245, 2020, doi: 10.1007/s41095-020-0179-3.

TABLE V

AUC SURFACE COVERAGE COMPARISON ACROSS METHODS. EACH EXPERIMENT IS REPEATED 10 TIMES AND RESULTS ARE AVERAGED OVER VIEWS 0-10 TO REDUCE VARIANCE ($\uparrow$). Δ : RELATIVE IMPROVEMENT OF OUR BEST METHOD OVER BEST COMPETING BASELINE. P : WELCH'S T-TEST ($N = 10$); AVAILABLE ONLY FOR OBJECTS WITH $\text{STD} > 0$. *** $P < 0.001$, ** $P < 0.01$, * $P < 0.05$, N.S. NOT SIGNIFICANT

Object	NBVNet [10]	PC-NBV [9]	RandomDiscrete	RandomContinuous	Our + 45°	Our + 22.5°	Our + 15°	Δ	p-value
nefertiti	0.441	0.780	0.573	0.563	0.924	0.755	0.845	+18.5%	$p = 0.353$ (n.s.)
suzanne	0.448	0.804	0.572	0.606	0.880	0.872	0.864	+9.5%	$p = 0.012^*$
igea	0.559	0.739	0.506	0.533	0.873	0.846	0.842	+18.1%	$p < 0.001^{***}$
cow	0.509	0.889	0.649	0.656	0.925	0.883	0.844	+4.0%	$p = 0.026^*$
xyzrgb_dragon	0.454	0.911	0.632	0.624	0.883	0.897	0.908	-0.3%	$p = 0.157$ (n.s.)
happy	0.352	0.845	0.581	0.589	0.834	0.818	0.840	+0.4%	$p = 0.564$ (n.s.)
teapot	0.482	0.836	0.681	0.663	0.745	0.810	0.740	-4.9%	$p = 0.118$ (n.s.)
stanford-bunny	0.435	0.816	0.594	0.560	0.906	0.823	0.874	+11.0%	$p = 0.029^*$
armadillo	0.419	0.860	0.615	0.602	0.894	0.804	0.820	+4.0%	$p = 0.224$ (n.s.)
CHILDREN_BEDROOM_CLASSIC	0.750	0.515	0.490	0.513	0.827	0.759	0.777	+10.3%	—
INTERNATIONAL_PAPER_Willamette	0.655	0.553	0.638	0.683	0.867	0.844	0.835	+26.9%	—
Melissa_Doug_Felt_Food_Pizza_Set	0.694	0.521	0.486	0.531	0.778	0.712	0.728	+12.1%	—
ASSORTED_VEGETABLE_SET	0.656	0.481	0.532	0.487	0.658	0.672	0.662	+2.4%	—
LEGO_Creationary_Game	0.834	0.748	0.703	0.730	0.915	0.893	0.883	+9.7%	—
MINI_ROLLER	0.620	0.574	0.526	0.490	0.635	0.554	0.614	+2.4%	—
Great_Dinos_Triceratops_Toy	0.716	0.764	0.646	0.695	0.866	0.797	0.770	+13.4%	—
PureConnect_2_Color_Fern	0.550	0.439	0.453	0.355	0.628	0.641	0.608	+16.5%	—

TABLE VI

CHAMFER DISTANCE COMPARISON ACROSS METHODS AT VIEW 10. EACH EXPERIMENT IS REPEATED 10 TIMES AND RESULTS ARE AVERAGED TO REDUCE VARIANCE ($\downarrow$). Δ : RELATIVE IMPROVEMENT OF OUR BEST METHOD OVER BEST COMPETING BASELINE. P : WELCH'S T-TEST ($N = 10$); AVAILABLE ONLY FOR OBJECTS WITH $\text{STD} > 0$. *** $P < 0.001$, ** $P < 0.01$, * $P < 0.05$, N.S. NOT SIGNIFICANT

Object	NBVNet [10]	PC-NBV [9]	RandomDiscrete	RandomContinuous	Our + 45°	Our + 22.5°	Our + 15°	Δ	p-value
nefertiti	0.00420	0.00278	0.00257	0.00317	0.00268	0.00292	0.00282	+31.9%	$p = 0.655$ (n.s.)
suzanne	0.00291	0.00206	0.00225	0.00211	0.00228	0.00245	0.00251	+2.4%	$p = 0.255$ (n.s.)
igea	0.00310	0.00299	0.00370	0.00366	0.00237	0.00255	0.00268	+20.7%	$p = 0.002^{**}$
cow	0.00209	0.00143	0.00177	0.00158	0.00178	0.00214	0.00227	-24.5%	$p < 0.001^{***}$
xyzrgb_dragon	0.00240	0.00144	0.00160	0.00165	0.00198	0.00179	0.00201	-24.3%	$p < 0.001^{***}$
happy	0.00399	0.00177	0.00231	0.00187	0.00207	0.00218	0.00225	-16.9%	$p < 0.001^{***}$
teapot	0.00252	0.00161	0.00171	0.00166	0.00275	0.00243	0.00399	-50.9%	$p = 0.024^*$
stanford-bunny	0.00523	0.00209	0.00223	0.00229	0.00233	0.00246	0.00256	+1.4%	$p = 0.077$ (n.s.)
armadillo	0.00351	0.00182	0.00196	0.00195	0.00219	0.00267	0.00256	-20.3%	$p = 0.002^{**}$
CHILDREN_BEDROOM_CLASSIC	0.00464	0.00628	0.00724	0.00671	0.00395	0.00412	0.00411	+14.9%	—
INTERNATIONAL_PAPER_Willamette	0.00774	0.00844	0.00666	0.00543	0.00506	0.00519	0.00520	+6.8%	—
Melissa_Doug_Felt_Food_Pizza_Set	0.00487	0.00594	0.00567	0.00572	0.00425	0.00450	0.00449	+12.7%	—
ASSORTED_VEGETABLE_SET	0.00492	0.00552	0.00535	0.00643	0.00464	0.00469	0.00486	+5.7%	—
LEGO_Creationary_Game	0.00328	0.00353	0.00374	0.00338	0.00267	0.00275	0.00282	+18.6%	—
MINI_ROLLER	0.00631	0.00478	0.00512	0.00541	0.00446	0.00467	0.00560	+6.7%	—
Great_Dinos_Triceratops_Toy	0.00404	0.00269	0.00303	0.00375	0.00218	0.00286	0.00312	+19.0%	—
PureConnect_2_Color_Fern	0.00654	0.00695	0.00756	0.00839	0.00482	0.00480	0.00520	+26.6%	—

- [14] A. Pop and L. Tamas, "Next best view estimation for volumetric information gain," *IFAC-PapersOnLine*, vol. 55, no. 15, pp. 160–165, 2022, doi: 10.1016/j.ifacol.2022.07.625.
- [15] J. I. Vasquez-Gomez, D. Troncoso, I. Becerra, E. Sucar, and R. Murrieta-Cid, "Next-best-view regression using a 3d convolutional neural network," *Machine Vision and Applications*, vol. 32, no. 2, p. 42, 2021, doi: 10.1007/s00138-020-01166-2.
- [16] S. Pan, H. Hu, H. Wei, N. Dengler, T. Zaenker, M. Dawood, and M. Bennewitz, "Integrating one-shot view planning with a single next-best view via long-tail multiview sampling," *IEEE Transactions on Robotics*, vol. 41, pp. 394–414, 2025, doi: 10.1109/TRO.2024.3507993.
- [17] P. Hoseini, S. K. Paul, M. Nicolescu, and M. Nicolescu, "A one-shot next best view system for active object recognition," *Applied Intelligence*, vol. 52, no. 5, pp. 5290–5309, 2022, doi: 10.1007/s10489-021-02657-z.
- [18] C. Collander, W. J. Beksi, and M. Huber, "Learning the next best view for 3d point clouds via topological features," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, p. 12207, doi: 10.1109/ICRA48506.2021.9561389.
- [19] G. Gao, M. Lauri, J. Zhang, and S. Frintrop, "Occlusion resistant object rotation regression from point cloud segments," in *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 2019, doi: 10.1007/978-3-030-11009-3_44.
- [20] S. Hinterstoisser, V. Lepetit, S. Ilic, S. Holzer, K. Konolige, G. Bradski, and N. Navab, "Technical demonstration on model based training, detection and pose estimation of texture-less 3d objects in heavily cluttered scenes," in *European Conference on Computer Vision*. Springer, 2012, p. 593, doi: 10.1007/978-3-642-33885-4_60.
- [21] S. Hinterstoisser, S. Holzer, C. Cagniat, S. Ilic, K. Konolige, N. Navab, and V. Lepetit, "Multimodal templates for real-time detection of texture-less objects in heavily cluttered scenes," in *2011 international conference on computer vision*. IEEE, 2011, pp. 858–865, doi: 10.1109/ICCV.2011.6126326.
- [22] G. Gao, M. Lauri, X. Hu, J. Zhang, and S. Frintrop, "Cloudaae: Learning 6d object pose regression with on-line data synthesis on point clouds," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, p. 11081, doi: 10.1109/ICRA48506.2021.9561475.
- [23] Y. Hu, P. Fua, W. Wang, and M. Salzmann, "Single-stage 6d object pose estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, p. 2927, doi: 10.1109/CVPR42600.2020.00300.
- [24] Z. Dong, S. Liu, T. Zhou, H. Cheng, L. Zeng, X. Yu, and H. Liu, "Ppr-net: point-wise pose regression network for instance segmentation and 6d pose estimation in bin-picking scenarios," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 1773–1780, doi: 10.1109/IROS40897.2019.8967895.
- [25] Y. Mei, S. Wang, Z. Li, J. Sun, and G. Wang, "Multi-modal 6-dof object pose tracking: integrating spatial cues with monocular rgb imagery," *International Journal of Machine Learning and Cybernetics*, vol. 16, no. 2, pp. 1327–1340, 2025, doi: 10.1007/s13042-024-02336-8.
- [26] Y. Su, M. Saleh, T. Fetzter, J. Rambach, N. Navab, B. Busam, D. Stricker, and F. Tombari, "Zebrapose: Coarse to fine surface encoding for 6dof object pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, p. 6728, doi: 10.1109/CVPR52688.2022.00662.
- [27] C. Wang, D. Xu, Y. Zhu, R. Martín-Martín, C. Lu, L. Fei-Fei, and

- S. Savarese, “Densefusion: 6d object pose estimation by iterative dense fusion,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, p. 3338, doi: 10.1109/CVPR.2019.00346.
- [28] M. Majcher and B. Kwalek, “Deep quaternion pose proposals for 6d object pose tracking,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 243–251, doi: 10.1109/ICCVW54120.2021.00032.
- [29] C. Papaioannidis and I. Pitas, “3d object pose estimation using multi-objective quaternion learning,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 8, pp. 2683–2693, 2020, doi: 10.1109/TCSVT.2019.2929600.
- [30] Y. Zhou, C. Barnes, J. Lu, J. Yang, and H. Li, “On the continuity of rotation representations in neural networks,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, p. 5738, doi: 10.1109/CVPR.2019.00589.
- [31] Wolfram, MathWorld. (2025) Rodrigues’ rotation formula. [Online]. Available: <https://mathworld.wolfram.com/RodriguesRotationFormula.html>
- [32] Q.-Y. Zhou, J. Park, and V. Koltun, “Open3D: A modern library for 3D data processing,” 2018, preprint.



Saulo A. Gante-Díaz received his degree in Control and Automation Engineering from the Escuela Superior de Ingeniería Mecánica y Eléctrica (ESIME) of the Instituto Politécnico Nacional (IPN) in 2020. In 2023, he obtained a Master in Computing Technology from the Centro de Innovación y Desarrollo Tecnológico en Cómputo of the IPN. He is currently pursuing a Doctorate degree in Computer Science at the Centro de Investigación en Computación of the IPN, Mexico City. His research interests include computer vision, robotics, artificial intelligence, neu-

ral rendering and 3D reconstruction.



Juan Irving Vasquez-Gomez received his MSc and PhD degrees from the National Institute for Astrophysics Optics and Electronics (INAOE), Mexico, in 2009 and 2014 respectively. He earned the B.S. degree in Computer Sciences from the Tehuacan Institute of Technology, Mexico, in 2006. Since February 2016, he is Researcher of the CONACYT assigned to the Instituto Politecnico Nacional (IPN) in Mexico. His research interests include robotics, motion planning, view planning and their applications to object reconstruction, inspection and surveil-

lance.



Erik Zamora is a full professor at the Instituto Politécnico Nacional (IPN). He earned a Diploma in electronics from UV (2004), followed by an M.Sc. in electrical engineering (2007) and a D.Sc. in automatic control (2015), both from CINVESTAV-IPN. He created the first commercial Mexican myoelectric prosthesis system and a robotic navigation system at the University of Bristol. He has authored more than 40 conference and journal papers and has supervised over 37 theses in these areas.