Grasp Pose Sampling for Precision Grasp Types with Multi-fingered Robotic Hands

Dimitrios Dimou[1] José Santos-Victor[1] Plinio Moreno[1]

***Abstract*— Generation of promising hand and finger poses for multi-fingered robotic hands cannot be simplified as the 2-dimensional model for grippers. Current approaches rely on heuristics that reduce the search space while ignoring a large number of candidates. We present a generative model that samples 6DoF poses for several types of precision grasps. Similarly to previous works, we start with a geometric heuristic to gather data. However, with a large enough samples we are able to sample grasp poses that are by a large margin more successful than using the heuristics. The model consists of 3 cascaded generative models that are based on the conditional Variational Auto-Encoder framework, and takes as input the desired grasp type, the object label, and the object's size. It generates a grasp posture, meaning the configuration of the fingers of the robotic hand, and a 6DoF pose. Our cascaded model samples first the finger joint configuration, followed by the Cartesian position of the object and finally the rotation of the object, our sampler divides the 6DoF in simpler problems, which lead to more successful grasps. In our experiments we show that our model improves the percentage of successful grasps sampled compared to the heuristic and compare several variants of the model to support our design choices, showing the benefits of the cascaded sampling.**

I. INTRODUCTION

Humans are able to grasp and handle objects with high competence even from a small age. That gives them the ability to directly manipulate their environment, use tools and cooperate efficiently [1]. As robots become ubiquitous in every-day life we will need them to be able to handle objects the same way that humans do, as most environments and objects are constructed in a way that facilitates the use from humans. Humans use different grasp types to grasp objects depending on the task to be performed, e.g. for handing over an object or for using it. Precision grasps are particularly used for handling small objects or performing precise movements [2], [3]. In precision grasp types the object is stabilized in the opposition created between the fingertips of two or more fingers. Inevitably dexterous robots should be able to perform precision grasps.

Grasping objects with dexterous robotic hands is a long standing research problem. Finding a finger configuration and a hand pose to achieve a specific grasp is extremely difficult, mainly due to the complexity that stems from the high number of degrees of freedom that anthropomorphic hands usually have. Ideally a grasp model should consider grasp types, which usually correspond to the type of task being executed. But modeling this kind of information in dexterous hands increases the complexity of the model. Recent advances on in-hand manipulation [4] have shown that low-level manipulation actions emerge by applying reinforcement learning in simulation. However, fine manipulation of objects of various shapes is still a very challenging problem, which is usually addressed by defining precision grasps.

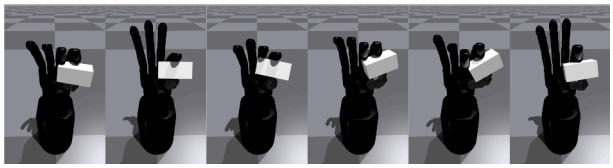

Fig. 1. Example grasps for each grasp type used in this work, executed with the Shadow Robot Hand. From left to right: tripod, palmar pinch, parallel extension, writing tripod, lateral tripod, and tip pinch.

To this end, we investigate the 1st step of this complex behaviour: sampling grasp poses for six precision grasp types using multi-fingered robotic hands. The goal of this work is to develop a grasp sampler that can generate the finger configuration and the hand pose for grasping simple shaped objects with a dexterous robotic hand using the following six precision grasp types: tripod, palmar pinch, parallel extension, writing tripod, lateral tripod, and tip pinch (Figure 1). Our model takes as input the grasp type that we want the robot to use, the 6DoF pose of the object and an estimation of the size of the object to be grasped and generates a grasp posture (i.e. the joint angles of the fingers) and a grasp pose (i.e. a three dimensional position and a rotation) for the hand base.

We follow a data-driven approach, in which we initially use a geometric heuristic to collect a large number of successful grasp examples in a simulated environment. Our heuristic is based on observations on the way humans tend to grasp objects using precision grasps, e.g. the location on the object from which it is grasped. We then use this dataset to train three conditional generative models: one to sample the finger configuration, one to sample the hand's position, and one to sample the hand's rotation. Each generative model is based on the conditional Variational Auto-Encoder (cVAE) framework. The model takes as input the grasp type that we want to perform, the object label, and an estimation of the grasp/object size and it generates a candidate object pose relative to the hand. Given then the 6DoF pose that the object is in we can compute a rigid transform for the hand base such that their relative pose remains fixed. In our experiments, we show the improvement in the percentage

[1]Institute for Systems and Robotics, Instituto Superior Tecnico, Universidade de Lisboa, Portugal. Emails: mijuomij@gmail.com, {jasv, plinio}@isr.tecnico.ulisboa.pt

of successful grasps generated by the model compared to the grasps computed using the heuristic. We also compare different variants of the model to show how our design choices affect the performance in grasp sampling.

In summary our contributions are the following:

- We present a geometric heuristic to generate candidate grasp poses for six precision grasp types. Our heuristic is based on observations on the way humans grasp, such as which fingers joints use for opposition and the position of the object.
- We develop a grasp pose sampling model, which generates promising precision grasps for a multi-fingered robotic hand. The grasp generation process is conditioned on the grasp type, the grasp size, the object label and the object size.
- We perform several ablation studies to support our design choices for each model.

II. RELATED WORK

Grasp sampling methods are usually divided in analytic or geometric methods and data-driven or learning-based approaches. Analytic methods require a model of the object and try to find contact points such that some grasp quality metric, like force closure, is optimized. These methods are based on accurate contact modeling, hand-crafted costs functions and time-consuming optimization processes. On the other hand, data-driven approaches collect datasets of successful grasps and train machine learning models to either: 1) approximate some success metric such as the probability of successfully grasping an object, 2) directly generate a candidate grasp using some kind of generative model, or 3) learn an end-to-end solution, like reinforcement learning where the output is a sequence of actions. For an overview of both analytic and learning-based approaches we refer the readers to the reviews [5], [6], [7]. In this work we follow a data-driven approach, so we focus our review on this category of methods.

In [8] they first generate a set of candidate grasp poses and then a network estimates the probability that each grasp for a given grasp type would be successful. Although this model generates grasp poses for multiple grasp types, the pose generation method was designed for a gripper and would be difficult to be adapted for human-like robotic hands. In [9] they present a network to generate dexterous grasps from point clouds. The network is trained with ground truth grasps by minimizing a consistency and a collision loss. After the network predicts a candidate grasp pose, a refinement process takes place where an optimization algorithm searches close to the proposed pose for a better candidate. In [10] they improved this model by adding a differentiable loss function that encodes a grasp quality metric and is optimized directly to find better grasp candidates. Because their model directly optimizes for each DoF of the hand, the resulting postures do not encode any grasp type information. In [11] they develop an architecture that evaluates the grasps proposed by a generative product of experts (PoE) model presented by [12], [13]. The PoE uses an object model, a contact model and a hand configuration model, while the evaluative architecture takes as input the grasp and a depth map of the scene and estimates the grasp success probability. Finally simulated annealing is performed to improve the grasp success probability. This model also does not generate grasps based on specific grasp types. In [14] they present a probabilistic graphical model to sample dexterous grasps which explicitly models the grasp type. They use maximum likelihood estimation to optimize the model on a set of successful grasps. This work can generate power and precision grasps but they model only one finger configuration for each grasp type. In addition they need a different grasp controller for each grasp type. In our work we focus only on precision grasp types but we model six different configurations and we control them using only one model.

Our work is inspired by the work presented in [15] where a cVAE is used to sample 6DoF grasp poses for a robotic gripper. They generated data by using a heuristic to estimate candidate grasp poses and simulated them in a physics engine. Using a generative model like the cVAE has the advantage that you can sample diverse poses that cover many possible ways that an object can be grasped. This approach was recently applied to multi-fingered hands in [16], [17], [18]. In [16], they train a cVAE, with successful grasps executed in simulation, to generate candidate grasps. The model was conditioned on the Basis Point Set encoding of a partial point cloud observation of the object. In addition, they train a grasp evaluator to predict the probability of success of each grasp, and use it to filter out low ranking grasps. In [17], a cVAE model is trained from grasps collected in simulation. They use point completion to complete the partial point cloud of an observed object, and then a PointNet architecture to extract a representation from the complete point cloud. This representation is used in the cVAE as a conditional variable. Finally they use a refinement procedure to optimize the contact points of the hand. In [18], a cVAE model, that is conditioned on the point cloud of object, is used to generate contact points on the objects surface. Then an optimization process computes the optimal finger joint angles to place the fingers on the generated contact points.

In our work, we propose a factorised model for generating candidate grasps. We use three cascaded cVAE models: one for generating the finger configuration, one for the hand position and one for the hand rotation. Each model is conditioned on the output of the previous model. This way we can generate multiple grasp poses for each finger configuration, in contrast to previous approaches that all the modalities where generated by one model. In addition, previous works where conditioning the cVAE models on some representation of the point cloud observation, we use the the label of the object and the size information. This way we can choose the side of the object that we want to grasp it from. Finally in our work we explicitly model the grasp type information as a conditional variable to the model. This way, during inference we can generate grasps of the grasp type we want to use.

III. BACKGROUND

In this work, we use a conditional Variational Auto-Encoder to model the grasping distributions. The cVAE consists of an encoder and a decoder network. The encoder takes as input a data point $\mathbf{x}$ and a corresponding conditional variable $\mathbf{c}$ and produces a latent point $\mathbf{z}$. The decoder takes as input a latent point $\mathbf{z}$ and a conditional variable $\mathbf{c}$ and generate a new data point $\mathbf{x}$. The encoder models the probability distribution $q(\mathbf{z}|\mathbf{x}, \mathbf{c})$, while the decoder the probability distribution $p(\mathbf{x}|\mathbf{z}, \mathbf{c})$. We train this model by maximizing the evidence lower bound (ELBO):

$$\mathcal{L}_{\boldsymbol{\theta},\boldsymbol{\phi}}(\mathbf{x}) = E_{q_\phi(\mathbf{z}|\mathbf{x},\mathbf{c})}\left[\log p_\theta(\mathbf{x}|\mathbf{c},\mathbf{z})\right] - D_{KL}\left(q_\phi(\mathbf{z}|\mathbf{x},\mathbf{c})\|p(\mathbf{z})\right)$$

The first term corresponds to the mean squared error between the reconstruction and the input, while the second minimizes the Kullback-Leibler divergence between the posterior of the model distribution $q(\mathbf{z}|\mathbf{x}, \mathbf{c})$ and the prior distribution $p(\mathbf{z})$.

IV. METHODS

The goal of this work is to develop a grasp sampler for simple shaped objects that can sample multiple grasp postures (i.e. finger joint angles) and 6DoF grasp poses (i.e. position and rotation of the hand) using precision grasp types. To achieve this we train three conditional generative models: a Posture Sampler, a Position Sampler, and a Rotation Sampler. The models are trained on a dataset of successfully executed grasps collected in a simulated environment. In this section we will first present the procedure that we used to generate our training data and the assumptions that our model is based on, and then we will present the details of the generative models.

Data generation process. Here we present the heuristic that is used to generate a candidate grasp pose for a given object and grasp type. The process is divided into three main parts: first a candidate grasp posture belonging to a specified grasp type is generated, second a candidate grasp pose is calculated, and third the grasp is executed and evaluated.

We assume that we already have a small dataset with 500 grasps, labeled with grasp type information for each grasp. With this dataset we train a initial cVAE model conditioned on the grasp type of each grasp which we will use to generate candidate grasps. This model is used only during the data generation procedure and later will be discarded. We use the following six precision grasp types, from the grasp taxonomy proposed in [2]: tripod, palmar pinch, parallel extension, writing tripod, lateral tripod, and tip pinch. These grasp types are achieved by stabilizing an object in the opposition that is defined between at least two joints, called opposition joints. This opposition is usually defined between joints in the thumb and the index finger and is called pad opposition [19]. The axis connecting these joints is called opposition axis. So for each grasp type a pair of opposition joints is determined. For the tripod, the writing tripod, the lateral tripod, and the tip pinch the opposition is created between the thumb tip and the index tip. For the the palmar pinch grasp, and the parallel extension the opposition is created between the thumb tip and the index middle. The first step then, is to generate a candidate grasp posture for a given grasp type with the initial cVAE and select the corresponding opposition joints. The second step is to calculate a candidate grasp pose. We break down this procedure into two phases: first we calculate the 3D position of the object and then the rotation. Instead of generating a candidate grasp pose as the position and rotation of the hand we assume that the hand is fixed in the origin of the coordinate axis and we generate a candidate pose for the object in the hand's reference frame. Based on the opposition joints selected for the given grasp type, we use forward kinematics to calculate the Cartesian positions of the rigid bodies of the opposition joints, when the grasp is executed. Then the object's position is calculated as the middle point between the two opposition rigid bodies. Various studies have found that humans grasp objects close to their center of mass for better stability [20], [21]. Finally, the rotation of the object is calculated such that one of the three canonical axis of the object is aligned with the opposition axis. We can see an example of a candidate grasp generation in Figure 2.

During the third step the object is placed in the calculated pose and the grasp is executed. The hand then performs a shaking movement to eliminate unstable grasp poses and the gravity is activated in the environment. If the object is still in the hand 5 seconds after gravity is activated the grasp is considered successful and the grasp type, the grasp posture, the grasp size, the object pose, the object type, and the object size is recorded. The grasp size is calculated as the distance (in cm) between the thumb tip and the index tip when executing the grasp without an object. The object size is calculated as the distance between the thumb tip and the index tip when executing the grasp with an object present. With this process we collect a large dataset of dexterous grasp postures executed on each object and using various grasp types.

Grasp sampler. Based on this dataset, the second goal of this work is to develop a generative model for sampling dexterous precision grasps given an object's pose, the object's type and the size of the side of the object that we want to grasp. More specifically we want to learn a sampler for the distribution of the successful grasps:

$$P(G \mid G_t, G_s, O_t, O_s)$$

where $G = (G_c, G_{pos}, G_{rot})$ is a successful grasp and consists of a grasp configuration G_c, representing the finger joint angles, and the three dimensional position of the hand G_{pos}, and the rotation G_{rot} for the hand base. The sampler is conditioned on the properties of the desired grasp: the type of the grasp G_t, and the size of the grasp G_s, and on the properties of the object that we want to grasp: the type of the object O_t, and the size of the object O_s.

Instead of directly modelling the full distribution:

$$P(G \mid G_t, G_s, O_t, O_s) = P(G_c, G_{pos}, G_{rot} \mid G_t, G_s, O_t, O_s)$$

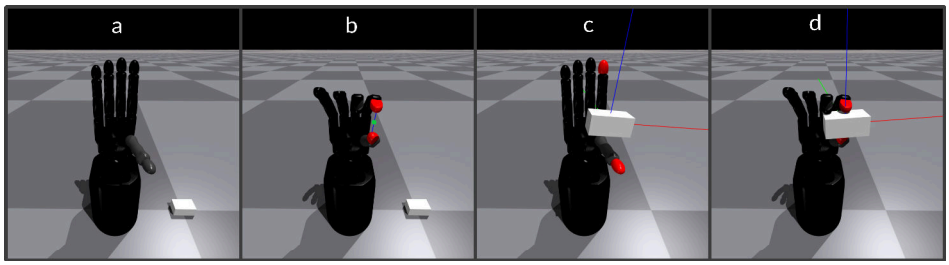

Fig. 2. Example for the process of generating a candidate grasp. **a**) The initial posture of the hand before grasping. **b**) A tripod a grasp posture is sampled from the initial cVAE model. The rigid bodies for the opposition joints can be seen in red. The blue line connecting them is the opposition axis and the green point is the middle point where the object will be placed. The grasp size recorded is the length of the blue line. **c**) The object is placed in the middle point and the blue axis of the object is aligned with the opposition axis. **d**) The grasp is executed, then the hand performs a shaking movement and gravity is activated. If the object remains in the hand the grasp is considered successful. The object size recorded is the current distance between the rigid bodies of the thumb and the index tip.

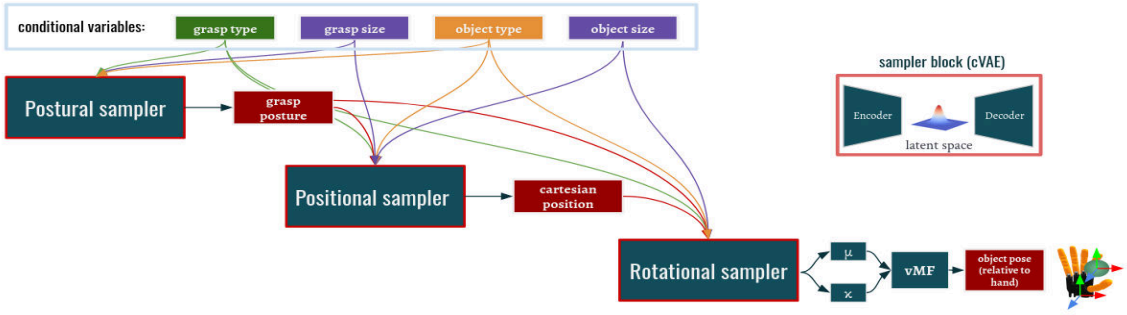

Fig. 3. Schematic representation of the Grasp Sampling procedure. The model consists of three individual samplers: the Postural Sampler that generates grasp postures (finger configurations), the Positional Sampler that generates the object's Cartesian position, and the Rotational Sampler that generates the object's rotation.

with one cVAE model as was done in previous works, we train three individual samplers: the Posture Sampler, the Position Sampler, and the Rotation Sampler, where each sampler is using a cVAE architecture. We show in the experiments that this approach leads to a model with higher performance. The Posture Sampler models the probability distribution:

$$P(G_c \mid G_t, G_s, O_t)$$

which is trained on all the successful grasp postures executed on the objects. It is conditioned on the grasp type label, the grasp size, and the object type. This model is based on the model for extracting conditional postural synergies, presented in [22]. The only difference is that in this work we also incorporate the grasp type as a conditional variable in the model. This way we can generate grasps of specific grasp types. The Position Sampler models the probability distribution:

$$P(G_{pos} \mid G_c, G_t, G_s, O_t, O_s)$$

and is trained on the 3D Cartesian positions of the objects. It is conditioned on the grasp posture, the grasp type label, the grasp size, the object type and the object size. Finally, the Rotation Sampler models the probability distribution:

$$P(G_{rot} \mid G_{pos}, G_c, G_t, G_s, O_t, O_s)$$

and is trained on the rotations of the objects, which are represented by quaternions. It is conditioned on the grasp posture, the grasp type label, the grasp size, the object type, the object size, and the Cartesian position. The Rotation sampler generates the mean direction and concentration parameter of a von Mises-Fisher distribution [23], which is then used to sample a rotation. We can see a schematic representation of the entire model in Figure 3. To obtain samples from the cascaded samplers, we provide the grasp type, the grasp size, the object type and the object size. Having sampled the 6DoF pose of the object, with the assumption that the hand is in the origin of the coordinate axis, we can calculate a pose of the hand for a given object pose that keeps their relative pose fixed. So given an object pose, a grasp type and a grasp size we can sample a hand pose to achieve this grasp.

V. Experimental Results

Set-up. In our experiments we use the IsaacGym simulator with the PhysX physics engine [24]. IsaacGym is a GPU accelerated simulator that can run multiple environments in parallel. We collected 270,000 grasps in total. The process took 10 hours on a NVIDIA GeForce RTX 3070. To train our models it takes around 20 minutes and to sample 1000 grasps 0.05 seconds. We used the Shadow Hand that has 24 joints. For the objects we used three boxes, three cylinders and three spheres of different sizes, which are depicted in Figure 4. The initial cVAE model, that generated candidate grasps in the dataset generation process, was trained on the dataset from [25], which contains 438 precision grasps with the

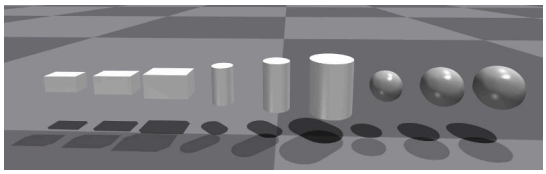

Fig. 4. The objects use in our experiments.

Shadow Hand that were collected by a human teleoperating the robotic hand through a data glove. It is conditioned on only the grasp type information without any information about the size of the grasp. For each grasp we record the grasp posture (i.e. the joint angles) in degrees, the grasp type, the grasp size (in cm), the object type, the object side size (in cm), the position and the rotation of the object. The grasp size for a given grasp posture is defined as the Euclidean distance between the thumb tip and the index tip when the grasp is executed without an object in the hand. The object size is also defined as the Euclidean distance between the thumb tip and the index tip but when the grasp is executed with an object in the hand.

Grasp Sampling Procedure. During test-time, to sample a new grasp we need the following information: 1) the grasp type that we want to perform which is one-hot encoded into a six dimensional vector, 2) the object type which is also one-hot encoded into a nine dimensional vector, 3) the actual size of the side of the object that we want to grasp it from. We use the actual size to compute the grasp size and the object size, as follows:

$$\text{grasp size} = \text{actual object side size} - 0.5\text{(cm)}$$

$$\text{object size} = \text{actual object side size} + 1.0\text{(cm)}$$

That is because when we record the object size we calculate the Euclidean distance between the thumb tip and the index tip which includes the size of the rigid bodies of the simulated robotic hand. The grasp size is reduced by a value of $0.5cm$, because we do not have any force feedback and we want the grasp to be firm enough to stably grasp the object.

Our model then generates: 1) the joint angles of the robotic hand, which are in degrees, 2) the position of the object with respect to the base of the hand, which is in centimeters and normalized to lie inside a unit sphere centered in the origin, 3) the rotation of the object, which is a unit quaternion. Given then the object's actual pose in the world frame we can compute the rigid transform of the hand such that the hand-object's relative transform remains fixed.

Evaluation metric. As an evaluation metric we used the percentage of successful grasps sampled and executed from each model. As our baseline we use the heuristic introduced in Section IV that generates object poses based on the opposition joints of each grasp. We present several variants of the grasp sampler model and compare them to support our choices on how we developed our final model. To compute the success rate for each model, we sample 500 grasps for each object and for each grasp type, totaling 27,000 grasps.

TABLE I

AVERAGE PERCENTAGE OF SUCCESSFUL GRASPS.

Model	Success Rate
Model 1	76.89
Model 2	79.41
Model 3	**90.90**
Heuristic	58.10

We execute the grasps, shake the hand, activate gravity and finally compute the percentage of successful grasps. We show the grasp success rate for all objects.

Quantitative Results. In our first experiment we investigate the advantage of using one model for each of the grasp posture, the object position and the object rotation, where each model is a cVAE sampler, compared to using one model to generate all the modalities as well as using two models: one for the grasp posture and one for the pose (position and rotation). All models take as input the grasp type, the grasp size, the object type, and the object side size. We can see the results of the average grasp success rate for each model in Table I.

Model 1 uses one cVAE sampler to represent all modalities, which models the distribution:

$$P(G_c, G_{pos}, G_{rot} \mid G_t, G_s, O_t, O_s)$$

Model 2 uses one cVAE for the grasp posture distribution, and one for the object pose distribution:

$$P(G_c \mid G_t, G_s, O_t, O_s)$$

$$P(G_{pos}, G_{rot} \mid G_c, G_t, G_s, O_t, O_s)$$

Model 3 uses one cVAE for the grasp posture distribution, one for the object position distribution, and one for the object rotation distribution:

$$P(G_c \mid G_t, G_s, O_t, O_s)$$

$$P(G_{pos} \mid G_c, G_t, G_s, O_t, O_s)$$

$$P(G_{rot} \mid G_{pos}, G_c, G_t, G_s, O_t, O_s)$$

In Table I we also see the success rates for the heuristic presented in Section IV. The model that uses one cVAE model for each modality outperforms all the other models. This suggests that factorising the grasp sampling and modelling each distribution separately describes our data better.

In our second experiment we investigate the utility of using the von Mises-Fisher distribution in our output layer of the rotation sampler network compared to having a non-linear layer and a normalization operation to transform the vector to a unit vector. The von Mises-Fisher (vMF) distribution is a distribution that allows to sample points on 4D spheres that are a natural representation for quaternions. We can see the results of the average grasp success rate for each model in Table II. The model which uses the vMF distribution to sample possible rotations outperforms the one with the neural network output layer.

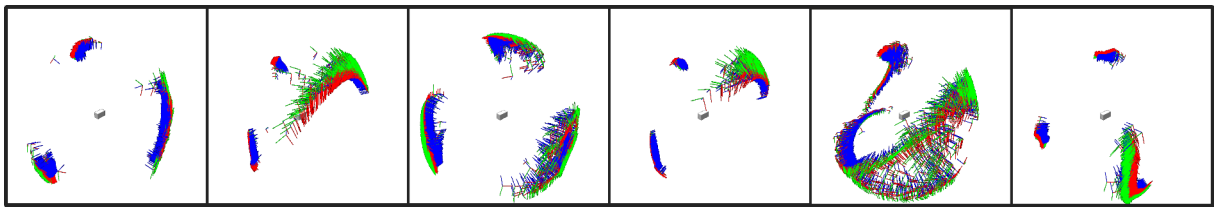

Fig. 5. The grasp poses collected during the data generation process for the medium box, for each grasp type. Each 6DoF grasp pose is represented by its coordinate frame, where each color represents a different axis.

TABLE II

AVERAGE PERCENTAGE OF SUCCESSFUL GRASPS.

Model	Success Rate
Model with NN	86.57
Model with vMF	**90.90**

TABLE III

CONDITIONAL VARIABLES USED FOR EACH MODEL VARIANT AND SUCCESS RATE.

	Position Sampler	Rotation Sampler	Success Rate
Variant 1	G_c, G_t, G_s, O_t, O_s	$G_{pos}, G_c, G_t, G_s, O_t, O_s$	90.90
Variant 2	G_c, G_t, G_s, O_t, O_s	$G_{pos}, G_c, G_t, O_t, O_s$	90.57
Variant 3	G_c, G_t, G_s, O_t	$G_{pos}, G_c, G_t, O_t, O_s$	90.36
Variant 4	G_c, G_t, O_t, O_s	$G_{pos}, G_c, G_t, O_t, O_s$	**91.54**
Variant 5	G_c, G_t, O_t	G_{pos}, G_c, G_t, O_t	81.68
Variant 6	G_c, G_s, O_t, O_s	$G_{pos}, G_c, G_s, O_t, O_s$	90.78
Variant 7	G_c, G_t, G_s, O_s	$G_{pos}, G_c, G_t, G_s, O_s$	71.77

Finally we investigate the effect of each conditional variable on the performance of the model. To this end, we trained the model with the three cVAE samplers using different subsets of the conditional variables for the Position and Rotation Samplers. The grasp posture sampler was kept the same in all model variants and it was conditioned on the grasp type, the grasp size and the object type. Table III shows the conditional variables used for the Position and Rotation Sampler for each model variant in the first and second column respectively and in the third column is the average grasp success rate for each model.

The model variants 1-4 show the effect of different combinations for the size variables. We notice that the results are similar for all models indicating that the size information that we will give to the model, either the grasp size or the object size, is not a significant factor to the model's performance. This makes sense since these two values are linearly dependent. The model variant 5 shows that if we completely remove the size information from both models then the performance decreases as well. The model variant 6 shows that removing the grasp type information does not affect much the performance of the model. This can be interpreted by the fact that the grasp position is mostly related to the finger tips of the hand which is encoded in the grasp posture variables, while the rotation is related with the size of the side of the object that we want to grasp. Finally, the model variant 7 does not encode any information about the object type, and is the worst performing model in terms of average successful sampled grasps, and highlights the importance of that parameter. In addition we can notice, that all models perform similarly on the sphere objects, where the rotation of the object does not affect the result since it is totally symmetric. Nevertheless, the model with the best performance is model variant 4 which uses the grasp and object type information, and the size of the object as conditional variables.

Qualitative Results. In Figure 5, we show the 6DoF grasp poses collected in the data generation process for the medium sized box for each grasp type. We see that each grasp type has a unique distribution of poses associated with it, that result from the shape of the object and the structure of the grasp type. In Figure 6, we see grasps sampled from our model for different objects. Each grasp is using a different grasp type. Our grasp sampling model could be easily applied in subsequent grasping tasks such as object pick-up. In that case additional steps would require to check for collisions for a sampled grasp pose, and plan the reaching of the end-effector to the specific pose which can be handled by trajectory optimization techniques.

Limitations and Future Work. Although our model is able to successfully generate precision grasps for a dexterous hand it produces only position commands (joint angles) without taking into account any force feedback thus is not able to adapt the grasp posture to uncertainties. Integrating additional modalities such as tactile feedback could potentially improve the application of our model in real-world scenarios. Finally, in this work we assume that all grasp types can be applied on each object. In reality humans use information related to the task that they want to perform in order to choose which grasp type they will use.

VI. CONCLUSION.

In summary, we presented a generative model that samples 6DoF grasp poses for six precision grasp types. We factorised the grasping process into three stages, and we trained a different model instance for each one. Our final model consists of three cascaded samplers: one for generating the finger configuration, one for generating the hand's position, and one for generating the hand's rotation. This way we are able to generate multiple grasp poses for each finger

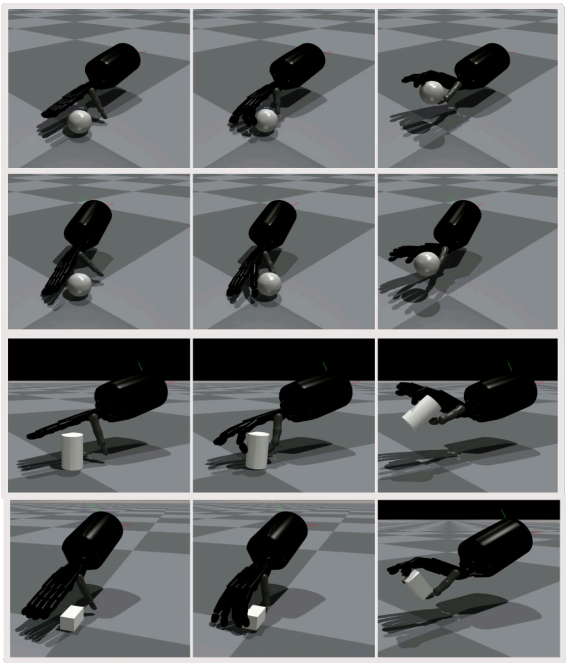

Fig. 6. Example grasps sampled from our model. Each row depicts a different grasp attempt. In the first column the hand is in the pose sampled by our model and the fingers in the pre-grasp position. In the second column the grasp is executed. In the third column the object is lifted to verify that the grasp is stable. The grasp types from the top row to bottom are the following: tripod grasp, palmar pinch, pinch, lateral tripod.

configuration. We also presented a geometric heuristic for calculating candidate grasp poses for each precision grasp, which we used to collect our dataset. Finally we demonstrated the benefits of the cascaded sampling approach in our experiments and supported our design choices based on quantitative results.

ACKNOWLEDGMENT

Work partially supported by the H2020 FET-Open project *Reconstructing the Past: Artificial Intelligence and Robotics Meet Cultural Heritage* (RePAIR) under EU grant agreement 964854, and by the Lisbon Ellis Unit (LUMLIS), and the FCT PhD grant [PD/BD/09714/2020].

REFERENCES

[1] A. Sobinov and S. J. Bensmaia, "The neural mechanisms of manual dexterity." *Nature reviews. Neuroscience*, 2021.

[2] T. Feix and R. Pawlik, "A comprehensive grasp taxonomy," 2009.

[3] T. Feix, J. Romero, H.-B. Schmiedmayer, A. M. Dollar, and D. Kragic, "The grasp taxonomy of human grasp types," *IEEE Transactions on Human-Machine Systems*, vol. 46, pp. 66–77, 2016.

[4] M. Andrychowicz, B. Baker, M. Chociej, R. Józefowicz, B. McGrew, J. W. Pachocki, A. Petron, M. Plappert, G. Powell, A. Ray, J. Schneider, S. Sidor, J. Tobin, P. Welinder, L. Weng, and W. Zaremba, "Learning dexterous in-hand manipulation," *The International Journal of Robotics Research*, vol. 39, pp. 20 – 3, 2020.

[5] A. Sahbani, S. El-Khoury, and P. Bidaud, "An overview of 3d object grasp synthesis algorithms," *Robotics Auton. Syst.*, vol. 60, pp. 326–336, 2012.

[6] J. Bohg, A. Morales, T. Asfour, and D. Kragic, "Data-driven grasp synthesis—a survey," *IEEE Transactions on Robotics*, vol. 30, pp. 289–309, 2014.

[7] R. Newbury, M. Gu, L. Chumbley, A. Mousavian, C. Eppner, J. Leitner, J. Bohg, A. Morales, T. Asfour, D. Kragic, D. Fox, and A. Cosgun, "Deep learning approaches to grasp synthesis: A review," *ArXiv*, vol. abs/2207.02556, 2022.

[8] M. Corsaro, S. Tellex, and G. D. Konidaris, "Learning to detect multimodal grasps for dexterous grasping in dense clutter," *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 4647–4653, 2021.

[9] M. Liu, Z. Pan, K. Xu, K. Ganguly, and D. Manocha, "Generating grasp poses for a high-dof gripper using neural networks," *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 1518–1525, 2019.

[10] ——, "Deep differentiable grasp planner for high-dof grippers," *ArXiv*, vol. abs/2002.01530, 2020.

[11] U. R. Aktas, C. Zhao, M. Kopicki, A. Leonardis, and J. L. Wyatt, "Deep dexterous grasping of novel objects from a single view," *ArXiv*, vol. abs/1908.04293, 2019.

[12] M. Kopicki, D. Belter, and J. L. Wyatt, "Learning better generative models for dexterous, single-view grasping of novel objects," *The International Journal of Robotics Research*, vol. 38, pp. 1246 – 1267, 2019.

[13] M. Kopicki, R. Detry, M. Adjigble, R. Stolkin, A. Leonardis, and J. L. Wyatt, "One-shot learning and generation of dexterous grasps for novel objects," *The International Journal of Robotics Research*, vol. 35, pp. 959 – 976, 2016.

[14] Q. Lu and T. Hermans, "Modeling grasp type improves learning-based grasp planning," *IEEE Robotics and Automation Letters*, vol. 4, pp. 784–791, 2019.

[15] A. Mousavian, C. Eppner, and D. Fox, "6-dof graspnet: Variational grasp generation for object manipulation," *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2901–2910, 2019.

[16] V. Mayer, Q. Feng, J. Deng, Y. Shi, Z. Chen, and A. Knoll, "Ffhnet: Generating multi-fingered robotic grasps for unknown objects in real-time," in *2022 International Conference on Robotics and Automation (ICRA)*, 2022, pp. 762–769.

[17] W. Wei, D. Li, P. Wang, Y. Li, W. Li, Y. Luo, and J. Zhong, "Dvgg: Deep variational grasp generation for dextrous manipulation," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 1659–1666, 2022.

[18] A. Wu, M. Guo, and C. K. Liu, "Learning diverse and physically feasible dexterous grasps with generative model and bilevel optimization," *ArXiv*, vol. abs/2207.00195, 2022.

[19] T. Iberall and C. L. MacKenzie, "Opposition space and human prehension," 1990.

[20] L. Desanghere and J. J. Marotta, "Graspability of objects affects gaze patterns during perception and action tasks," *Experimental Brain Research*, vol. 212, pp. 177–187, 2011.

[21] S. L. Prime and J. J. Marotta, "Gaze strategies during visually-guided versus memory-guided grasping," *Experimental Brain Research*, vol. 225, pp. 291–305, 2012.

[22] D. Dimou, J. Santos-Victor, and P. Moreno, "Learning conditional postural synergies for dexterous hands: A generative approach based on variational auto-encoders and conditioned on object size and category," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, 2021, pp. 4710–4716.

[23] S. Sra, "Directional statistics in machine learning: A brief review suvrit sra," *Applied Directional Statistics*, 2018.

[24] V. Makoviychuk, L. Wawrzyniak, Y. Guo, M. Lu, K. Storey, M. Macklin, D. Hoeller, N. Rudin, A. Allshire, A. Handa, and G. State, "Isaac gym: High performance gpu-based physics simulation for robot learning," *ArXiv*, vol. abs/2108.10470, 2021.

[25] A. Bernardino, M. Henriques, N. Hendrich, and J. Zhang, "Precision grasp synergies for dexterous robotic hands," in *2013 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, 2013, pp. 62–67.