Variational GIGA: Variational Grasp Detection via Implicit Representations

Zhiyao Bao[†], Zhenyu Jiang[†], Yuke Zhu[†]

[†]The University of Texas at Austin

***Abstract*—Variational grasp detection seeks to generate various grasps through a generative model. In this work, the robots need to analyze the objects in a 3D scene (either in pile scenario or packed scenario as shown in Figure 1), and then learn grasps that try to remove all objects in the scene. In the context of the original GIGA [9], which harnesses the synergies between affordance and geometry, we propose this work to extend the regression model to a generative model. Our model takes advantage of the generative model to predict the full distribution of viable grasp parameters to generate various grasps for one grasp center through an encoder-latent space-decoder network structure. Same as GIGA, we train the model on self-supervised grasps trials data in simulation, generating one ground-truth grasp as the label for each grasp center. The evaluation is performed on simulated grasping tasks. With objects laid on a 3D scene, the robots need to remove these objects in clutter by grasping one object every trial. Variational GIGA is evaluated by both the variability and the quality of the generated grasps. We only performed experiments in simulation. The experiments demonstrate the great variability generated through our model and the good quality of predicted grasps (which is at the same level as the original GIGA). Besides, through analyzing the results, we discovered that there is a trade-off between the diversity and quality of the trained model.**

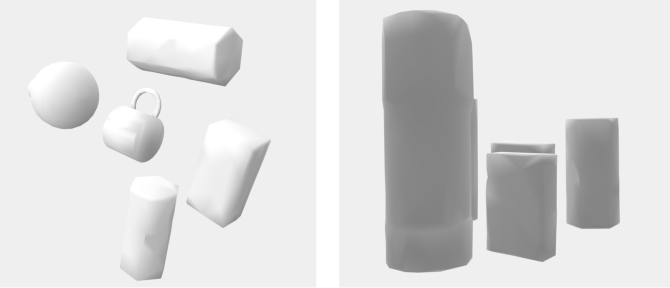

Fig. 1: We perform the task of clutter removal. These are examples of visualization in simulation of two scenarios. In the pile scenario (left), objects are dropped with random positions and poses. In the packed scenario (right), objects are in the canonical 3D poses, placed upright and ordered like pieces of stuff being placed by individuals in daily life. Our model will reason about the 3D scene like these and generate reconstructed grasps to remove one object at a time until all the objects in the 3D scene are removed.

I. Introduction

Generating variational robust grasps is very important. One reason is it is a very prevalent task using grippers to physically interact with the environment to grasp objects. Thus, improving the diversity and quality of grasps is critical. Another reason is with variational grasps, the model is able to increase its exploration ability, investigating more possible grasping poses.

This task requires the robots to reason about the 3D scene and infer various reconstructed grasps' parameters from the 3D scene. We formalize the problem in the context of removing objects in clutter through 6-DoF grasp detection with a single side-view camera. The main object is generating a reasonable amount of reconstructed grasps for each grasp center on the scene of a clutter of objects so that the grippers can grasp and remove one object at a time from the scene to declutter all.

Inspired by the idea of implicit geometry and affordance representations and variational grasp generation, we investigate if a generative model will improve grasping performance. our initial thought is if there is a good grasp reconstructed for a certain grasp center, there will be more good grasps for that grasp center. In this case, generating more grasp per grasp center may maximize the benefits of implicit geometry and affordance representations.

The original GIGA [9] utilizes the synergy between 3D reconstruction and grasp affordance for 6-DoF grasp detection in a clutter removal task. However, it only predicts one ground truth grasp pose for each grasp center. We are wondering if predicting the full distribution of grasp parameters would improve grasp detection. Therefore, we introduced our work, Variational GIGA (Variational Grasp Detection via Implicit Geometry and Affordance). It extends the grasp affordance part to learn to predict the full distribution of viable grasp parameters in a generative model rather than a single ground truth grasp pose for each grasp center. It is approached through a network structure similar to cVAE. Encoder predicts the mean and standard deviation of the distribution, which then are used to generate the latent features. The decoder predicts the grasp parameters from the latent features. Through modifying the network structure, we are able to achieve a generative model with the prediction of the full distribution of viable grasp parameters.

There are two main challenges during the whole process. First, the primary challenge is to develop a model that is able to generate multiple grasps based on the grasp center and local features. We need to figure out a way to modify the regression model to a generative one. We applied a network structure similar to cVAE to turn it into a generative model, predicting the distribution of grasp parameters rather than directly predicting grasp parameters to solve this issue. Second, after changing to a generative model, more hyperparameters (e.g. the weight of KL-Divergence loss) are added. It may require some fine-tuning. We tune the model using the small dataset and the large dataset to solve it, which will be explained in detail later.

We did experiments in simulation. Either in the pile or packed scenario, objects are placed in the 3D scene and our model generates grasp parameters to perform clutter removal tasks. Objects are removed one at a time by the gripper (based on the generated grasp parameters). The quality of the generated grasps is on the same level compared to the original GIGA. First, The validation accuracy is on the same level compared to GIGA. Training with the large dataset, in pile scenario Variational GIGA reached 88.3% validation accuracy while in packed scenario Variational GIGA reached 90.07% validation accuracy. Second, the Grasp Success Rate and Declutter Rate results are at the same level as the original GIGA. Evaluating the model trained with the large dataset, in pile scenario it reached 63.22% Grasp Success Rate and 50.57% Declutter Rate in the pile scenario, and 77.6% Grasp Success Rate and 77.2% Declutter Rate in the packed scenario. Besides evaluating the quality of grasps, our model has more variability and diversity compared to the original GIGA as shown in Figure 5. It not only is able to generate more grasps per grasp center, but also may generate more grasps on different grasp centers overall as it explores more.

We summarize the main contributions of our work as follows:

- We introduce a generative model for affordance prediction that effectively predicts the distribution of grasp parameters.
- We provide a detailed analysis of experiments results. Comparing our model's results to the results from the original GIGA can clearly show the advantages and disadvantages of our model. We also derive the trade-off between diversity and quality for variational grasp detection.

II. Related Work

Our work is primarily based on GIGA [9]. As the next generation of GIGA, most of the architecture of the network and the datasets are the same as GIGA.

A. Implicit Geometry and Affordance

Implicit representations have been improved and used a lot in recent years ([4], [12], [8]). GIGA [9], which applies implicit representations of geometry and affordance, takes the benefits of differentiable and continuous implicit functions.

Since our model only wants to predict more grasps per grasp center, the implicit representations are still effective in our case. Thus, our model maintains similar implicit representations as the original GIGA to take the benefits from implicit functions. While keeping the geometry implicit function the same as the original GIGA, the affordance implicit function is modified to a structure similar to cVAE since we would like to predict the full distribution of grasp parameters. The affordance and geometry are still jointly learned simultaneously to harness the synergies between them.

B. Variational Grasp Detection

There are pioneer works of the variational grasp generation ([7], [11]).

The theory of the evidence lower-bound objective (ELBO) [7] is the basis theoretically showing the validity of variational auto-encoder (VAE), which maximizes ELBO when training. As one of the primary methods to get generative models, it has been applied to perform several tasks like handwritten digits generation, sentence prediction, sentence interpolation and so on. A variant of VAE is the conditional variational auto-encoder (cVAE), which generates a conditional distribution.

As a state-of-the-art method for getting generative models, we applied cVAE to affordance prediction, conditioning on each location which is based on the corresponding 3D local features. Our model's network structure of affordance prediction is very similar to the variational grasp sampler [11]. Both works have a latent space satisfying the Gaussian distribution, which is chosen beforehand. The detailed implementations are a little bit different. The backbone of our model is the convolutional occupancy network [13] while the related work [11] uses PointNet++ [14]. Our encoder takes in the grasp center and its corresponding local features while that related work [11] takes in the point cloud and a grasp. Our decoder generates the qualities and rotations of the reconstructed grasps while that related work [11] generates the grasp poses. Overall, the high-level idea of variational grasp detection is very similar to those related work while the low-level implementations vary.

III. Data

A. Datasets

We used the dataset from the GIGA [9] directly. And they originally come from VGN [3]. It is not only because these data are good enough for us to use, but also because using the same data may provide a fair comparison between our model (Variational GIGA) and the original GIGA.

The dataset is a synthetic grasping dataset generated with physics simulation. The object sets can be found through this link.

The ground-truth grasp labels are collected in simulation. Grippers will generate physical trials during the self-supervised data generation process, so that the ground truth quality and rotation labels are able to be collected. One ground-truth grasp label is collected for each grasp center. The occupancy labels are collected from object meshes. After the self-supervised data generation process, there is a process for data cleaning, data balancing, and noise adding to make

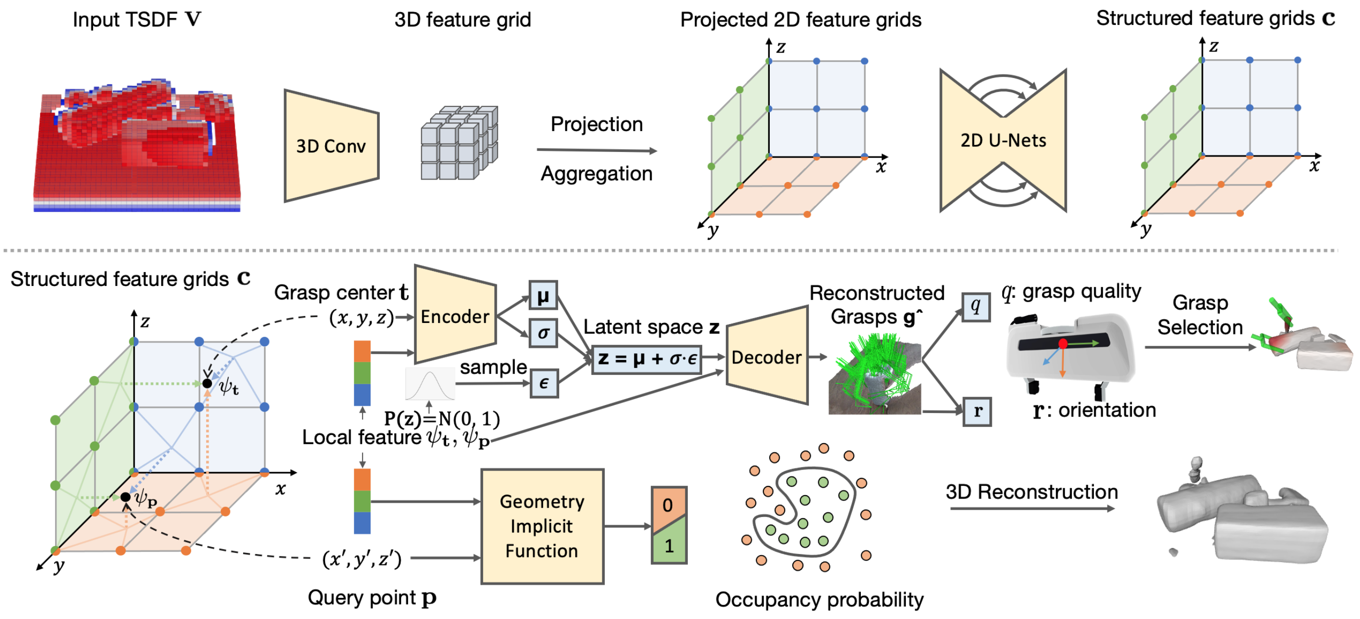
Input TSDF $\mathbf{V}$
3D Conv
3D feature grid
Projection
Aggregation
Projected 2D feature grids
z
x
y
2D U-Nets
Structured feature grids $\mathbf{c}$
z
x
y
Structured feature grids $\mathbf{c}$
z
x
y
ψ_t
ψ_p
Grasp center $\mathbf{t}$
(x, y, z)
Encoder
μ
σ
ϵ
sample
P(z)=N(0, 1)
Local feature ψ_t, ψ_p
Latent space $\mathbf{z}$
$\mathbf{z} = \mu + \sigma \cdot \epsilon$
Decoder
Reconstructed Grasps $\hat{\mathbf{g}}$
q
$\mathbf{r}$
q: grasp quality
$\mathbf{r}$: orientation
Grasp Selection
Geometry Implicit Function
(x', y', z')
Query point $\mathbf{p}$
0
1
Occupancy probability
3D Reconstruction

Fig. 2: The model architecture of Variational GIGA, which learns the affordance distribution rather than a single point regression. Compared to the GIGA model architecture, we changed the affordance implicit functions into cVAE, the variational autoencoder. The Encoder Network takes in the grasp center and the corresponding local features, and outputs the mean and standard deviation of the probability density function of latent space. Along with a random variable sampled from Normal Distribution, we predict the latent space. The Decoder Network takes in latent features and local features, and outputs the reconstructed grasps that will be decomposed to grasp parameters. Another minor change is we removed width in the output since it is useless when generating grasps.

sure these self-generated data are of good quality. These data preprocessing and data generation processes are the same as the original GIGA.

For training, we generate two datasets for each scenario. The small dataset contains around 10,000 data points for each scenario. The large dataset contains around 10M and 2M data points for pile and packed scenarios respectively.

B. Simulation Environment

Our simulation environment is also the same as the original GIGA since it is good to use and it is easy to do a fair comparison when running experiments. The simulation environment is built on Pybullet [5]. Same as the original GIGA, our model uses a free gripper to sample grasps in a 30 × 30 × 30 cm^3 tabletop workspace. There are two simulated scenes, pile and packed as shown in Figure 1. When running experiments, we use 5 random seeds with 50 simulation runs (simulated grasps) each for evaluation.

IV. Problem Formulation

We consider the problem of 6-DoF grasp detection for unknown rigid objects in clutter from a single-view depth image. Currently, we are curious if extending GIGA to learn the affordance distribution rather than a single point regression will improve grasp learning.

A. Assumptions

The robot arm is a parallel-jaw gripper. The workspace is initialized with several unknown rigid objects. The model takes in a Truncated Signed Distance Function (TSDF) generated by a single-view camera. The model outputs the 6-DoF grasp pose predictions and grasp qualities. [9] Also, we assume the probability density function of the latent space is in Gaussian Distribution.

B. Notations

Grasps We define a 6-DoF grasp g as the grasp center position $\mathbf{t} \in \mathbb{R}^3$, and the orientation $\mathbf{r} \in SO(3)$ of the gripper. [9] The opening width $w \in \mathbb{R}$ between the fingers is removed from the original GIGA since it is useless in grasping.

Grasp Quality A scalar grasp quality $q \in [0, 1]$ estimates the probability of grasp success. The grasp quality q is one-dimensional. We learn to predict the grasp quality of a grasp with binary success labels of executing the grasp trial in simulation. [9]

Occupancy For an arbitrary point $\mathbf{p} \in \mathbb{R}^3$, the occupancy $b \in \{0, 1\}$ is a binary value indicating whether this point is occupied by any of the objects in the scene. [9]

C. Objectives

Goal The primary goal for Variational GIGA is to detect 6-DoF grasps through the synergies between 3D reconstruction

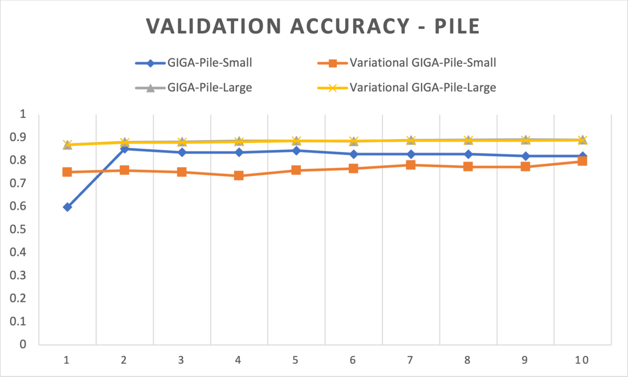

Fig. 3: We log the validation accuracy (average validation quality of generated grasps) using a line chart. The x-axis is the number of epochs while the y-axis is the accuracy generated in the pile scenario. The small dataset has around 10,000 data points for training. The large dataset has around 10M data points for training. GIGA-Pile-Small is the validation accuracy generated by training the original GIGA with a small dataset in the pile scenario. GIGA-Pile-Large is the validation accuracy generated by training the original GIGA with a large dataset in the pile scenario. Variational GIGA-Pile-Small is the validation accuracy generated by training our current model with a small dataset in the pile scenario. Variational GIGA-Pile-Large is the validation accuracy generated by training our current model with a large dataset in the pile scenario.

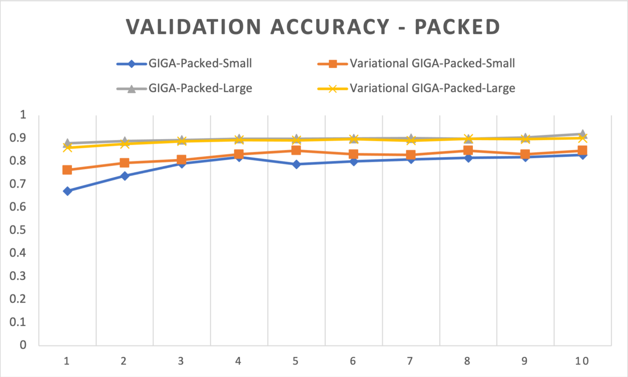

Fig. 4: Same as Figure 3, we log the validation accuracy (average validation quality of generated grasps) using a line chart. The y-axis is the accuracy generated in the packed scenario.

The small dataset has around 10,000 data points for training. The large dataset has around 2M data points for training. There are fewer data points in the packed scenario than the pile scenario, we originally trained with 20 epochs rather than 10, but our current model (Variational GIGA) converges at around 10 epochs, so for a fair comparison, we trained for 10 epochs for all. The labels in this figure have the same meaning as Figure 3 except everything is in the packed scenario.

and *variational* grasp generation. Given the input observation $\mathbf{V}$, our goal is to learn three functions:

$$\begin{aligned} f_e &: \mathbf{t} \rightarrow \mu, \sigma, \\ f_d &: \mathbf{z} \rightarrow q, \mathbf{r}, \\ f_g &: \mathbf{p} \rightarrow b. \end{aligned} \tag{1}$$

The first function f_e maps from a grasp center point $\mathbf{t} \in \mathbb{R}^3$ and its corresponding local features $\phi_{\mathbf{t}}$, $\phi_{\mathbf{p}}$ to the mean μ and standard deviation σ of the probability density function of latent space. Once f_e is trained, we can obtain the best latent space probability distribution.

The second function f_d maps from a latent feature $\mathbf{z}$ to rotation r and grasp quality q of the best grasp. Once f_d is trained, we can select which grasp to execute based on the grasp quality at grasp centers.

We use f_e and f_d combined to learn the posterior distribution P(G* | $\mathbf{V}$), where G* is the reconstructed grasp space and $\mathbf{V}$ is the input observation. Each grasp g $\in$ G* can be decomposed to r and q, which are rotation and quality of the grasp respectively. While predicting μ and σ through f_e, we generate a random variable ϵ by sampling from N(0, 1), the Normal Distribution. The latent feature is calculated by $\mathbf{z}$ = μ + $\sigma \cdot \epsilon$, which is input into f_d to generate the reconstructed grasp. [1]

The third function f_g maps any point in the workspace to the estimated occupancy value at that point. We can extract a 3D mesh from the learned occupancy function with the Marching Cube algorithm [10].

Expected Results And Evaluation We will take the experiment results of GIGA as the baseline to compare. We will evaluate based on the validation set's grasp quality during training, and grasp success rates (GSR) and declutter rates (DR) during testing. We will also visualize the grasp affordance landscape and predicted grasps to evaluate. We expect the new method with a full distribution of grasp pose parameters will improve grasp detection exploration, increasing the variability or diversity of grasps generated. It is expected that variational GIGA will generate more various grasps compared to GIGA.

V. Method

We now present Variational GIGA, a model that leverages synergies between 3D reconstruction and *variational* affordance for 6-DoF grasp detection from partial observation.

Besides cVAE, there are a lot of other approaches to reach the goal of a generative model like the Gaussian Mixture Model, Probabilistic Context-free Grammar, and so on. Some do not fit into our grasp detection case, while others may not meet our need to get a generative model conditioning on each location of the object based on its corresponding local features. And the intuitive approach is using cVAE, which directly solves the problem and reaches the goal.

The grasp affordance learns the full distribution of viable grasp parameters. Figure 2 illustrates the overall model architecture. We no longer have a single network, the affordance implicit functions to predict affordance parameters through a single point regression. Rather, we use a structure similar to cVAE [2], which includes an encoder network and a decoder network, to predict grasp parameters. By applying the idea of variational autoencoder, we are able to get a generative model, which is exactly what we want for generating multiple grasps per grasp center. In this case, our cVAE network structure can successfully approach our goal. The details of the modified parts are discussed as follows.

A. Encoder

Our encoder takes in the grasp center $\mathbf{t}$ and local features $\phi_{\mathbf{t}}$, $\phi_{\mathbf{p}}$ as inputs and processes them to generate the mean μ and standard deviation σ of the distribution of the latent space. We will tune the dimension of latent size with 128 and 64.

B. Latent Space

We use reparameterization tricks to generate the latent feature. Basically, $\mathbf{z} = \mu + \sigma \cdot \epsilon$, where μ and σ are outputs of the Encoder, ϵ is sampled from normal distribution, and $\mathbf{z}$ is the input of the Decoder. Since $\epsilon \sim \mathbf{N}(0, 1)$, the latent space satisfies the Gaussian distribution.

C. Decoder

Our decoder takes in the local features $\phi_{\mathbf{t}}$, $\phi_{\mathbf{p}}$ and latent features $\mathbf{z}$ as inputs, and processes them to generate the grasp quality q and grasp orientation $\mathbf{r}$. The output dimension of grasp quality q is 1 while output dimension of grasp orientation $\mathbf{r}$ is 3.

D. Loss Function

According to the variational lower-bound (ELBO) theory [7], we need to add the Kullback-Leibler divergence loss (KL-divergence loss) into the original loss function. The calculation of KL-divergence loss is based on the VAE tutorial [6]. We will tune the weight of KL-divergence loss in the loss function with 1, 0.1 and 0.01.

The overall loss function is:

$$L = \frac{1}{n} \cdot \sum_{0}^{n} L(\hat{q}, q) + \frac{1}{n} \cdot \sum_{0}^{n} L(\hat{\mathbf{r}}, \mathbf{r}) + L(\hat{b}, b) + \omega \cdot KLDivLoss$$

Here L represents the loss function, n represents the number of grasps generated per grasp center, $\hat{q}$ represents the predicted grasp quality, q represents the ground-truth grasp quality, $\hat{\mathbf{r}}$ represents the predicted grasp rotation, $\mathbf{r}$ represents the ground-truth grasp rotation, $\hat{b}$ represents the predicted occupancy, b represents the ground-truth occupancy, ω represents the weight of KL-Divergence loss (which is a hyperparameter to tune), and KLDivLoss represents the KL-Divergence Loss calculated through the torch.nn.functional.kl_div() Pytorch internal function.

For each grasp center, n grasps are generated. We average the affordance loss of those n grasps of the same grasp center so that the affordance loss and the geometry loss are on the same scale.

VI. Experiments and Results.

We study the efficacy of variational grasp generation through experiments. Through the experiments, we are trying to investigate into one question: we are wondering if variational GIGA is able to improve grasping, evaluating based on diversity/variation and accuracy/quality of the grasps generated. We use the results from the original GIGA [9] as the baseline.

A. Training Results

We first trained models with the small dataset and the large dataset. We trained 10 epochs for all. Considering the large dataset of the packed scenario has only 2M data, we originally trained the large dataset in the packed scenario with 20 epochs to compensate for that. However, our current model converges at 10 epochs so we end up training all with 10 epochs. We set the latent size to be 64, the weight of KL-Divergence loss to be 0.1, and the number of grasps generated per grasp center to be 10 in our model. All other hyperparameters (e.g. learning rate) are the same as the original GIGA.

The validation accuracies (validation quality of generated grasps) are shown in Figure 3 and 4 for pile and packed scenarios respectively. Training with the large dataset, the validation accuracy of our model achieved 88.83% in the pile scenario and 90.07% in the packed scenario. According to Figure 3 and 4, we can find that when training with the large dataset, GIGA and Variational GIGA has similar performance according to the validation accuracy. The validation accuracy curve varies for the small dataset. Considering the size of the small dataset, the variance may be large. Thus we may only conclude that overall the validation accuracies generated from our model are approximately at the same level compared to the original GIGA.

B. Grasp Detection Results

We evaluate our model on two perspectives: 1) diversity (or variation) of the generated grasps; 2) quality (or accuracy) of the generated grasps.

1) Grasp Diversity Analysis:

As our model predicts variational grasps, evaluating the diversity of grasps generated is essential. It is not only because it may help to show the validity of our implementations, but also because it illustrates the exploration of grasps predicted.

We take the original GIGA as the baseline [9]. We tried to evaluate the diversity quantitatively. However, since the original GIGA only has 1 predicted grasp per grasp center, our model has better diversity on each grasp center for sure. The original GIGA may not provide a quantitative baseline for us to compare. Therefore, we tried to evaluate the diversity qualitatively, mainly through visualization. Because we removed width entirely in our model since it is useless, it is still needed in the visualization to show the width of the gripper. In visualization, we use the ground-truth width of

Method		Pile		Packed	
		GSR (%)	DR (%)	GSR (%)	DR (%)
Small Datatset	GIGA	16.75 ± 4.48	8.91 ± 2.88	24.54 ± 9.03	20.51 ± 9.30
	Variational GIGA	16.28 ± 1.46	8.92 ± 1.20	25.56 ± 5.41	21.05 ± 5.41
Large Dataset	GIGA	68.78 ± 4.23	49.73 ± 4.08	79.44 ± 2.35	78.94 ± 3.00
	Variational GIGA	63.22 ± 1.57	50.57 ± 3.02	77.60 ± 6.54	77.20 ± 6.32

TABLE I: Quantitative Experiments Results. We tested both the original GIGA and our model (Variational GIGA) on clutter removal tasks through simulated grasps. The mean and standard deviation of GSR (Grasp Success Rate) and DR (Declutter Rate) of each scenario are shown.

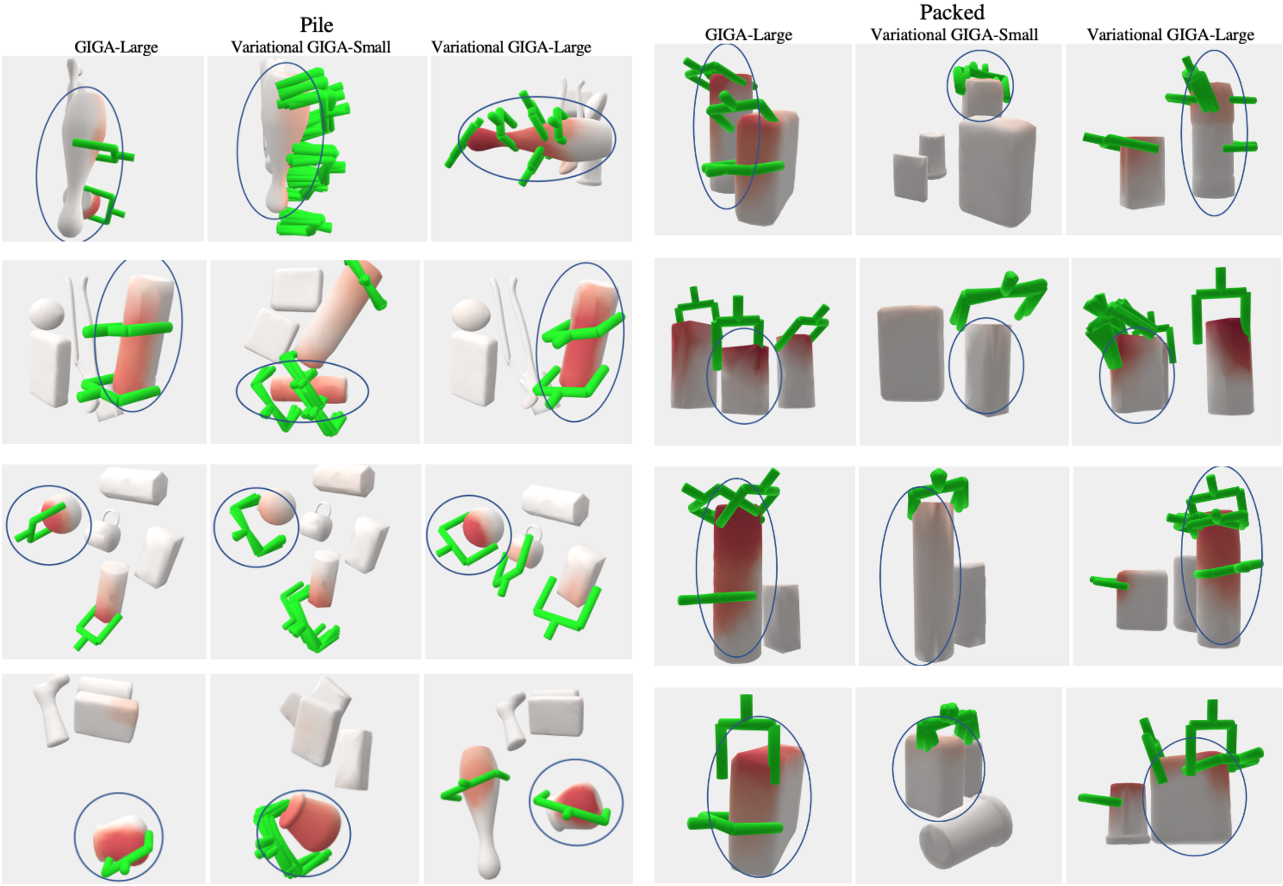

Fig. 5: Grasp visualization for pile (left) and packed (right) scenarios. The blue circles indicate the objects of interest.

simulated grasps (which is the width of the gripper when it contacts the objects).

The visualization of generated grasps is shown in Figure 5. For each scenario, the first column illustrates the visualization of the original GIGA training with the large dataset, the second column illustrates the visualization of our model training with the small dataset, and the third column illustrates the visualization of our model training with the large dataset. Because we generate simulated grasps with random seeds and objects in the scene are not always removed at the grasp trial, the visualization we generated on the same row does not have the same set of objects in the 3D scene. However, it does not affect the analysis of diversity in our case since we are able to discern certain objects of interest to compare. The blue circle points out the object of interest in the scene.

In the pile scenario of Figure 2, we show the visualization of a bowling-pin-shaped object, an irregular-top cylindrical object, a spherical object, and a trapezoidal-column-shaped object from row 1 to row 4 respectively. The model trained with the small dataset is able to generate a great number of various grasps for a grasp center. The model trained with the large dataset doesn't generate a lot of grasps per grasp center, but it may generate more grasps in different grasp centers overall (e.g. the grasps generated on the bowling-pin-shaped object). We believe that is because our model can explore more grasp poses for a single grasp center and thus has a higher probability to predict one with good grasp quality.

In the packed scenario of Figure 2, we show the visualization of an irregular-top cylindrical object, a cuboidal object, a cylindrical object, and a Rounded cuboidal object from row 1 to row 4 respectively. Same as the pile scenario, the model trained with the small dataset can generate a great number of various grasps for a grasp center. The model trained with the large dataset is able to generate multiple grasps per grasp center for the first three objects.

In conclusion, our model is able to generate more diverse grasps both in the pile and packed scenarios compared to the original GIGA.

2) *Grasp Quality Analysis:*

The quality of grasps generated by our model is also important. We report the Grasp Success Rate (GSR) and Declutter Rate (DR) in Table I. We trained the original GIGA with the same dataset and the same hyperparameters. We evaluated based on the same object set (pile/test and packed/test) on the simulated grasp task. Using the same metric as the original GIGA, we are able to conduct a fair comparison between our model and the baseline. Our model achieved 63.22% GSR and 50.57% DR in the pile scenario, and 77.6% GSR and 77.2% DR in the packed scenario. The numbers are slightly lower than the original GIGA, but they are still on the same level compared to GIGA. Thus, we can conclude Variational GIGA is able to generate grasps of approximately the same level of quality as the original GIGA.

3) *Trade-off Between Diversity and Quality:*

After obtaining and analyzing the results of our experiments, we find there is a trade-off between diversity and quality in our model.

The evidence for this trade-off is explained as follows. First, we find that the model trained with the small dataset is of higher diversity but worse quality, while the model trained with the large dataset is of lower diversity but better quality. Second, The model trained with the large dataset in the packed scenario generates more diverse grasps compared to the model trained with the large dataset in the pile scenario. Given the fact that the large dataset of pile scenario has 10M data while the large dataset of packed scenario has 2M data, the packed model is expected to generate more diverse grasps than the pile model.

In summary, Variational GIGA achieved the same level of grasp quality and better grasp diversity compared to the original GIGA. It is able to achieve the goal of learning to predict the full distribution of viable grasp parameters with generative modeling. The trade-off between diversity and quality in Variational GIGA exists. Training with a larger dataset approaches better quality while training with a smaller dataset approaches better diversity. In applications, balancing between the grasp quality and the grasp diversity may be needed.

VII. Conclusion

We introduced Variational GIGA, which is able to perform 6 DoF grasp detection in clutter removal tasks. As the next generation of GIGA [9], our model learns variational grasp detection by generative modeling rather than generating a single grasp per grasp center. We approached the generative model through a network structure similar to cVAE. It is able to predict the full distribution of viable grasp parameters for each grasp center, which is the most important advantage of our work. We did experiments in simulation. Through experiments, we investigated the grasp diversity and quality of our model. We conclude that Variational GIGA can achieve better grasp diversity and the same level grasp quality compared to the original GIGA. We also find that the trade-off between grasp diversity and quality exists in our model, which is a possible weakness and needs to be concerned in applications. Another weakness of our work is we did not test our model on real robots. However, all the results we obtained demonstrate that overall Variational GIGA outperforms the original GIGA.

There are two possible future extensions of our work. First, we plan to generate more ground-truth grasp poses for each grasp center in the data generation process. Thus, it may further improve the diversity of grasps. Second, we hope to use more complex grippers. Currently we are using the parallel-jaw gripper. We are considering if using more complex grippers in Variational GIGA may improve grasp quality.

References

[1] Dieter Fox Arsalan Mousavian, Clemens Eppner. 6-dof graspnet: Variational grasp generation for object manipulation. *arXiv:1905.10520v2*, 2019.

[2] Tim Baumgärtner. Variational autoencoder conditional variational autoenoder on mnist. https://github.com/timbmg/VAE-CVAE-MNIST.git, 2018-2019.

[3] Michel Breyer, Jen Jen Chung, Lionel Ott, Siegwart Roland, and Nieto Juan. Volumetric grasping network: Real-time 6 dof grasp detection in clutter. In *Conference on Robot Learning*, 2020.

[4] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5939–5948, 2019.

[5] Erwin Coumans and Yunfei Bai. Pybullet, a python module for physics simulation for games, robotics and machine learning. http://pybullet.org, 2016–2019.

[6] Carl Doersch. Tutorial on variational autoencoders, 2021.

[7] Kuan Fang, Yuke Zhu, Animesh Garg, Silvio Savarese, and Li Fei-Fei. Dynamics learning with cascaded variational inference for multi-step manipulation. *Conference on Robot Learning (CoRL)*, 2019.

[8] Moritz Ibing, Isaak Lim, and Leif Kobbelt. 3d shape generation with grid-based implicit functions, 2021.

[9] Zhenyu Jiang, Yifeng Zhu, Maxwell Svetlik, Kuan Fang, and Yuke Zhu. Synergies between affordance and geometry: 6-dof grasp detection via implicit representations. *Robotics: science and systems*, 2021.

[10] William E Lorensen and Harvey E Cline. Marching cubes: A high resolution 3d surface construction algorithm. *ACM Siggraph Computer Graphics*, 21(4):163–169, 1987.

[11] Arsalan Mousavian, Clemens Eppner, and Dieter Fox. 6-dof graspnet: Variational grasp generation for object manipulation, 2019.

[12] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[13] Songyou Peng, Michael Niemeyer, Lars Mescheder, Marc Pollefeys, and Andreas Geiger. Convolutional occupancy networks. *arXiv preprint arXiv:2003.04618*, 2020.

[14] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017.