Supplementary Material of Implicit Grasp Diffusion

1 Preliminary: Diffusion Models

Diffusion models aim to model an unknown data distribution $q(\boldsymbol{x}_0)$ with a parameterized model $p_\theta(\boldsymbol{x}_0)$. The procedure consists of two steps: the forward and the reverse diffusion processes. The forward process iteratively injects small Gaussian noise in $\boldsymbol{x}_0$ to obtain $\boldsymbol{x}_{1:T}$:

$$q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0) = \prod_{t=1}^{T} q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}), \tag{1}$$

where $q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) = \mathcal{N}(\boldsymbol{x}_t; \sqrt{1-\beta_t}\boldsymbol{x}_{t-1}, \beta_t \boldsymbol{I})$ is the per-step noise injection following variance schedule $\beta_1, ..., \beta_T$. This leads to the distribution $q(\boldsymbol{x}_t|\boldsymbol{x}_0) = \mathcal{N}(\boldsymbol{x}_t; \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0, (1-\bar{\alpha}_t)\boldsymbol{I})$, where $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$. Since $\bar{\alpha}_t \approx 0$, $\boldsymbol{x}_T \sim \mathcal{N}(0, \boldsymbol{I})$. The reverse diffusion learns to denoise the data starting from $\boldsymbol{x}_T$ following $p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) = \mathcal{N}(\boldsymbol{x}_{t-1}; \mu_\theta(\boldsymbol{x}_t, t), \beta_t \boldsymbol{I})$ where:

$$\mu_\theta(\boldsymbol{x}_t, t) = \frac{1}{\sqrt{\alpha_t}}(\boldsymbol{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t)). \tag{2}$$

The parameterized model $\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t)$ is called the score function, and it is trained to predict the perturbations and the noising schedule by the score-matching objective:

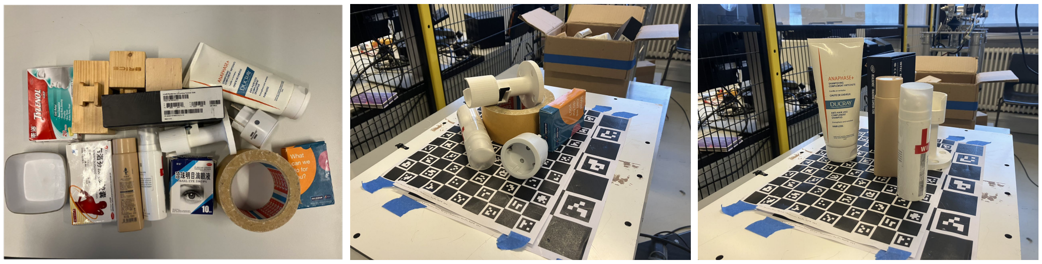
$$\arg\min_\theta \mathbb{E}_{t\sim[1,T], \boldsymbol{x}_0 \sim q, \boldsymbol{\epsilon} \sim \mathcal{N}(0,\boldsymbol{I})} \left[||\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t)||^2 \right]. \tag{3}$$

In particular, such a score function represents the gradient of the learned probability distribution as:

$$\nabla_{\boldsymbol{x}_t} \log p_\theta(\boldsymbol{x}_t) = -\frac{1}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t). \tag{4}$$

2 Experiment Scenes

The objects used in the real-world experiment are shown in Fig. 1a. Fig. 1b and Fig. 1c illustrate examples of packed and pile scenes.



(a) (b) (c)

Figure 1: (a) Objects for the real-world declutter experiment. (b) An illustration of the pile scene. (c) An illustration of the packed scene.

8th Conference on Robot Learning (CoRL 2024), Munich, Germany.

Table 1: Robustness experiments. We evaluate the performance under four types of noises: dex, stereo, lateral, blur, and depth. N denotes sampling rounds in IGD. We set dex noise as the baseline to calculate the performance decrease (in bracket). The best performances are highlighted in red.

Noise Type	Method	Packed		Pile	
		GSR (%)	DR (%)	GSR (%)	DR (%)
Dex (baseline)	GIGA [2]	84.8±2.2	85.1±2.5	69.5±1.3	49.0±3.4
	IGD (N=1)	92.9±1.8	86.7±1.8	68.2±1.9	50.6±1.5
	IGD (N=9)	91.2±0.9	88.8±1.5	71.0±0.7	55.0±1.6
Stereo	GIGA [2]	72.0±1.6 (↓ 12.8)	78.4±2.6 (↓ 6.7)	51.8±1.2 (↓ 17.7)	47.0±2.3 (↓ 2.0)
	IGD (N=1)	83.6±1.8 (↓ 9.3)	86.4±2.1 (↓ 0.3)	60.4±0.9 (↓ 7.8)	48.6±1.6 (↓ 2.0)
	IGD (N=9)	85.6±1.3 (↓ 5.6)	88.3±1.2 (↓ 0.5)	61.9±1.5 (↓ 9.1)	54.8±2.5 (↓ 0.2)
Lateral	GIGA [2]	75.1±1.4 (↓ 9.7)	81.5±1.0 (↓ 3.6)	56.3±1.5 (↓ 13.2)	53.8±2.8 (↑ 4.8)
	IGD (N=1)	83.7±1.2 (↓ 9.2)	86.4±1.8 (↓ 0.3)	63.6±1.5 (↓ 4.6)	53.2±2.4 (↑ 2.6)
	IGD (N=9)	85.2±1.6 (↓ 6.0)	87.2±1.4 (↓ 1.6)	68.6±2.9 (↓ 2.4)	61.2±2.2 (↑ 6.2)
Blur	GIGA [2]	69.6±2.4 (↓ 15.2)	72.4±1.2 (↓ 12.7)	53.8±1.9 (↓ 15.7)	48.2±2.7 (↓ 0.8)
	IGD (N=1)	81.9±1.7 (↓ 11.0)	79.8±2.2 (↓ 12.7)	60.9±2.9 (↓ 7.3)	43.5±3.1 (↓ 7.1)
	IGD (N=9)	84.2±1.5 (↓ 7.0)	84.5±1.7 (↓ 4.3)	63.4±2.4 (↓ 7.6)	48.4±3.0 (↓ 6.6)
Depth	GIGA [2]	75.3±1.7 (↓ 9.5)	83.6±1.2 (↓ 1.5)	51.8±1.8 (↓ 17.7)	47.6±2.1 (↓ 1.4)
	IGD (N=1)	89.9±0.6 (↓ 3.0)	88.6±1.4 (↑ 1.9)	58.5±0.8 (↓ 9.7)	49.2±1.4 (↓ 1.4)
	IGD (N=9)	90.0±1.1 (↓ 1.2)	90.1±0.5 (↑ 1.3)	62.1±1.3 (↓ 8.9)	56.5±3.1 (↑ 1.5)

3 Visualization of Grasp Detection

We visualize the top-10-score grasps with a threshold of 0.5 in some challenging cases in Fig. 3. Compared to GIGA, IGD generates more collision-free good-quality grasps and gives lower scores to bad-quality grasps.

4 Robustness to Different Noises

Table 1 shows the performance of GIGA and IGD in different kinds of noises. According to [1], noise in a depth camera includes the following noises: noise from stereo matching, lateral noise, blur, and noise in depth estimation. We simulate those noises and add them in the depth image to evaluate the performance of models. Our baseline condition is the dex noise environment, where models are trained. According to Table 1, both IGD(N=1) and IGD(N=9) outperform GIGA in different noise conditions. Besides, the performance decrease of IGD is also lower than GIGA. The results demonstrate the strong robustness of IGD. The robustness of IGD comes from two factors: (i) Probabilistic two-stage grasp evaluator can precisely estimate the true quality of grasps; (ii) Multi-round sampling increases the chance to sample good grasps. From Table 1, increasing sampling rounds decreases performance drop. Due to its strong robustness, IGD performs well in real-world experiments.

5 Ablation Studies

We also conducted extensive ablation studies to validate each module in our proposed IGD. Ablation studies were mainly performed in pile scenes because of data abundance and higher task difficulty compared to packed scenes. Table 2 shows the ablation studies of IGD. The proposed DAM effectively improves the performance in both GSR and DR. Besides, the performance deteriorates with GC alone compared to the AE. Combining the AE and GC achieves the best performance, which demonstrates the effectiveness of the proposed probabilistic two-stage grasp evaluator. To further investigate what brings the performance improvement, we train AE and GC together and deactivate one of them in the inference (shown as ”GC*” and ”AE*” in Table 2). If we deactivate the GC, the performance decreases drastically, while the performance decline from deactivating the AE is limited. This result is contrary to solely training the AE and GC. We can conclude that the performance improvement is mainly from the loss supervision instead of a simple ensemble of two modules. Negative grasp sampling is also important to train the grasp evaluator. Introducing neg-

Table 2: Ablation Studies of the proposed IGD. ”DAM” denotes the deformable attention module. ”AE” denotes the affordance evaluator. ”GC” denotes the grasp classifier. ”Neg.” denotes the negative grasp sampling for the grasp classifier. ”AE*” and ”GC*” denote that AE and GC are trained together, but they are used solely in the inference. The best performances are highlighted in red.

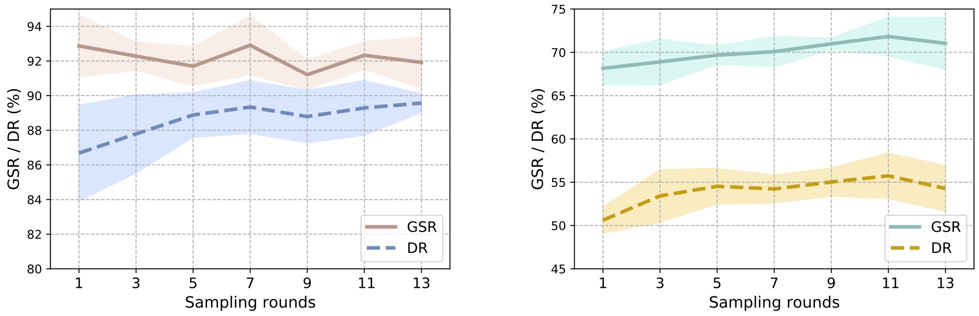
DAM	AE	GC	Neg.	AE*	GC*	GSR (%)	DR (%)
	✓					59.1±2.8	42.8±3.1
✓	✓					62.2±3.0	45.6±2.5
✓		✓				53.5±2.6	42.5±4.0
✓		✓	✓			72.2±1.8	33.5±2.0
✓	✓	✓				62.8±3.1	47.9±4.3
✓	✓	✓	✓			68.2±1.9	50.6±1.5
✓			✓	✓		59.5±0.9	47.5±0.9
✓			✓		✓	64.3±2.3	48.0±3.1

Table 3: Ablation Studies of anchor points in DAM. The best performances are highlighted in red.

Anchor points	GSR (%)	DR (%)
2^3	68.2±1.9	50.2±1.5
3^3	65.7±2.3	49.0±2.7
4^4	62.3±1.9	46.2±2.5

Table 4: Ablation Studies of focal loss. The best performances are highlighted in red.

γ	α	GSR (%)	DR (%)
0	0.5	68.5±2.9	46.1±3.9
1	0.25	65.7±1.9	43.6±1.4
1	0.5	66.5±1.8	49.7±1.4
1	0.75	63.2±2.6	47.5±2.6
2	0.25	66.3±2.1	46.9±3.1
2	0.5	68.2±1.9	50.6±1.5
2	0.75	65.2±3.5	50.4±2.9
3	0.5	62.7±3.5	45.8±3.3



(a)

(b)

Figure 2: The ablation study of sampling rounds. (a) Packed scene. (b) Pile scene.

ative grasp sampling with GC alone largely increases GSR (53.5% to 72.2%) but decreases DR (42.5% to 33.5%). When it comes to probabilistic two-stage structure, negative sampling improves both GSR and DR. The results mean that the grasp evaluator needs to excavate information from negative samples to learn to distinguish feasible grasps in the large grasp space.

Number of sampling rounds. Since we can sample multiple rounds to obtain different grasps at the same grasp position, an ablation study about sampling rounds is conducted to analyze the effect of this hyper-parameter, which is shown in Fig. 2. In packed scenes, there is no improvement in GSR as sampling rounds increase, while DR increases in general. In contrast, in harder pile scenes, both

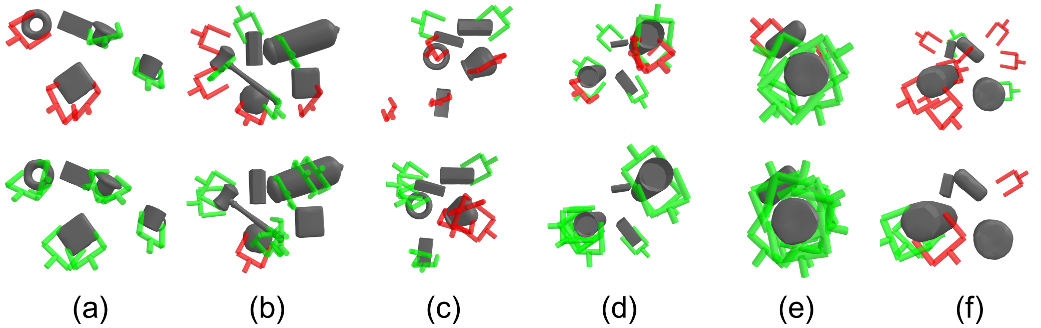


Figure 3: Grasp visualization in some challenging cases. The first row denotes GIGA, and the second row denotes IGD. (a-c) are in pile scenes, while (d-f) illustrate packed scenes. Green grasps denote successful grasps, while red grasps denote failed grasps.

GSR and DR increase as sampling rounds increase. The results show the ability of IGD to obtain good grasps by increasing grasp samples, which is inherited from sampling-based methods.

Anchor points in DAM. Table 3 shows an ablation study on anchor points in DAM. Because the anchor points are initialized as the points spatial-uniformly sampled in the cube around the grasp center, we select 2^3, 3^3, and 4^3 anchor point numbers to evaluate their performance. From Table 3, the best performance appears in the 2^3 setting.

Focal loss hyper-parameters. Table 3 shows an ablation study of focal loss hyper-parameters in GC. $\gamma = 0$ reduces the focal loss to a normal cross-entropy loss and shows inferior performance. γ is to tune the level of hard example mining, and α is to balance the weight of positive and negative samples. According to the results, we select the best hyper-parameter setting as $\alpha = 0.5$ and $\gamma = 2$.

6 Limitations and Future Work

Although we have demonstrated the effectiveness of IGD in both simulation and real-world systems, there are limitations that future work can improve. First, we can't directly obtain the point with the highest affordance value. In GIGA, $40 \times 40 \times 40$ points are sampled to query the grasp to obtain the best grasp. IGD inherits this limitation from GIGA. Second, diffusion models have higher computational costs and inference latency compared to dense prediction methods, especially in IGD where grasp sampling is separated into two stages: position sampling and orientation sampling. Future work can exploit the latest advancements in diffusion model acceleration methods to reduce the number of inference steps required, such as new noisy schedules [3], inference solvers [4], and consistency models [5]. Third, in our Grasp Diffuser, we directly apply diffusion models to the generation of quaternions. However, quaternion in the diffusion process is not in close form. Current works have proposed a lot of methods to achieve rotation diffusion [6, 7, 8]. Although our effort to apply these works to IGD fails, it is still worth exploring.

References

[1] T. Mallick, P. P. Das, and A. K. Majumdar. Characterizations of noise in kinect depth images: A review. *IEEE Sensors journal*, 14(6):1731–1740, 2014.

[2] Z. Jiang, Y. Zhu, M. Svetlik, K. Fang, and Y. Zhu. Synergies between affordance and geometry: 6-dof grasp detection via implicit representations. *arXiv preprint arXiv:2104.01542*, 2021.

[3] T. Chen. On the importance of noise scheduling for diffusion models. *arXiv preprint arXiv:2301.10972*, 2023.

[4] C. Lu, Y. Zhou, F. Bao, J. Chen, C. Li, and J. Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Advances in Neural Information Processing Systems*, 35:5775–5787, 2022.

[5] Y. Song, P. Dhariwal, M. Chen, and I. Sutskever. Consistency models. *arXiv preprint arXiv:2303.01469*, 2023.

[6] J. Urain, N. Funk, J. Peters, and G. Chalvatzaki. Se (3)-diffusionfields: Learning smooth cost functions for joint grasp and motion optimization through diffusion. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5923–5930. IEEE, 2023.

[7] A. Leach, S. M. Schmon, M. T. Degiacomi, and C. G. Willcocks. Denoising diffusion probabilistic models on so (3) for rotational alignment. In *ICLR 2022 Workshop on Geometrical and Topological Representation Learning*, 2022.

[8] Y. Jagvaral, F. Lanusse, and R. Mandelbaum. Unified framework for diffusion generative models in so (3): applications in computer vision and astrophysics. *arXiv preprint arXiv:2312.11707*, 2023.