Supplementary Material for "Improving Robotic Grasp Detection Under Sparse Annotations via Grasp Transformer with Pixel-wise Contrastive Learning"

May 21, 2024

In this document, we include a detailed experimental setup, comprehensive physical and real-world experiments, extensive ablation studies for each module, and intuitive visualization analysis to help readers better understand our study and demonstrate the superiority of our proposed sparse detection framework GT-PCL.

1 Experimental Setup

Following the methodology established by previous works such as GraspNet [1], GGCNN [2], and GK-Net [3], our approach begins with the utilization of a normalized image as the primary input to the model. We aim to predict three pixel-wise maps, each sharing the same resolution as the input image, and for consistency, the batch size is set to the default value of 16. This design choice aligns with the prevailing practices in the field and enables a standardized comparison with state-of-the-art methods.

The core of our model comprises a Transformer encoder, which plays a pivotal role in understanding the image-based representations of the grasping task. In Table 1, we provide a comprehensive overview of all relevant configurations, encompassing critical parameters such as the head number denoted as N and the feature channels specified as C. The configurations are meticulously selected to optimize the model's performance and are fine-tuned in empirical experiments.

The decoder model, inspired by the work of Kimura et al. [4], follows a convolutional architecture. This design choice facilitates the translation of the rich feature representations learned by the Transformer encoder into meaningful grasp predictions. The seamless integration of the encoder and decoder components contributes to the model's ability to generate precise grasp maps.

To train our model effectively, we adopt the Huber loss function, which has proven to be highly suitable for regression tasks such as grasp prediction. This loss function strikes a balance between Mean Squared Error (MSE) and Mean Absolute Error (MAE), making it robust to outliers in the data. During the training phase, we optimize the model using the Adam optimizer, a widely used choice for deep learning tasks. The default learning rate is set at 1e-4, though it can be fine-tuned based on specific requirements and experiments.

All our training and evaluation processes are carried out using Nvidia RTX 1080Ti GPUs, known for their computational efficiency in deep learning workloads. The physical system hosting the computations runs on the Ubuntu 18.04 platform with an Intel Core i7

Table 1: Training configuration for each layer in GT-PCL.

GT-PCL	Output Size	Layer Name	Configuration
Stage1	56×56	Multi-scale Attention Block	$r_i = \begin{cases} 4 & i < \frac{head}{2}, \\ 8 & i \geq \frac{head}{2}. \end{cases}$ $C_1 = 64, N_1 = 2, head_1 = 4$
Stage2	28×28	Multi-scale Attention Block	$r_i = \begin{cases} 2 & i < \frac{head}{2}, \\ 4 & i \geq \frac{head}{2}. \end{cases}$ $C_2 = 128, N_2 = 2, head_2 = 8$
Stage3	112×112	Transposed Conv Block	$C_3 = 128, W_3 = 4, stride_3 = 4$
Stage4	224×224	Transposed Conv Block	$C_4 = 64, W_4 = 2, stride_4 = 2$

processor, ensuring a stable and well-supported environment for our experiments. In the interest of providing comprehensive and reliable results, we conducted each experiment ten times and reported the averaged results. This meticulous approach minimizes the impact of random variations and strengthens the validity of our findings, ensuring a fair comparison with existing methods in the field.

2 Ablation for Shunted Attention

A key adaptation is the introduction of the grouped-query technique into the shunted attention mechanism. As depicted in Figure 1, original shunted attention [5] employs a shared single Q-Projector to generate query keys (Q) for dot-product attention operations with diverse key-value pairs (K-V). While such implementation simplifies the computational process via using a shared network as the Q-Projector, it still encounters limitations in capturing multi-granularity information for effective grasp detection in clustered scenarios. Intuitively, a shared Q Projector inevitably leads the model to favor features of a specific granularity, while generating queries at the head level introduces excessive computational load. A similar idea is also mentioned in [6]. Therefore, adopting specific query heads for various granularities is to achieve an effective trade-off between performance and efficiency. Compared to single-query shunted attention, we fabricate the shunted attention mechanism via incorporating the grouped-query Q-Projectors, each designated to generate corresponding queries for respective granularity groups within all shunted K-V pairs. This simple yet efficient modification enables the model to extract semantic information from a broader spectrum, which is crucial for understanding complex scenarios involving objects of multiple scales. Experimentally, we supplemented a quantitative analysis on two commonly used datasets, Cornell [7] and Jacquard [8]. To provide a fair evaluation, we adopt single-stage supervised training *from scratch* under both dense and sparse annotation settings. The results are presented in Table 2. The grouped-query shunted attention can consistently achieve better grasp detection accuracy on two benchmarks, offering a stronger baseline for sparse grasping scenarios. Notably, there remains a certain performance gap compared to our proposed

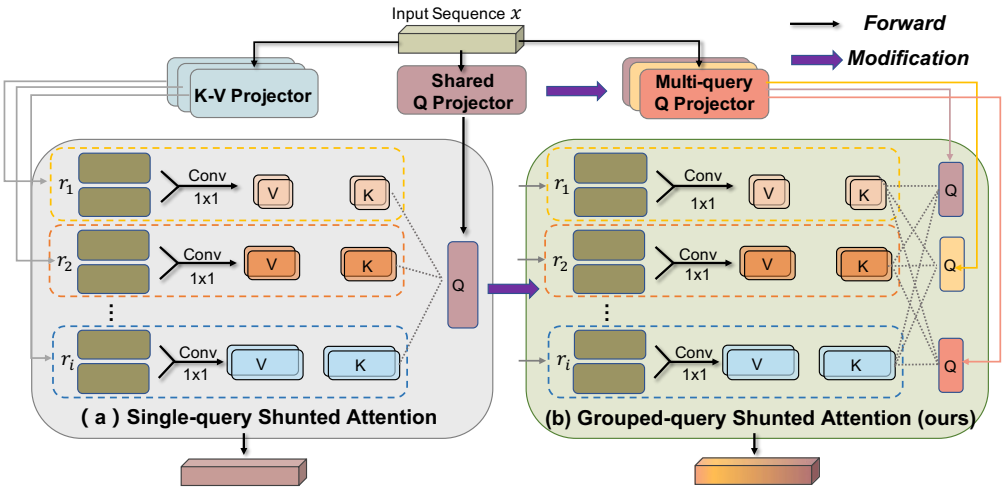

Figure 1: The Comparison of **Single-query Shunted Attention** (left) and **Grouped-query Shunted Attention** (right). The black arrows represent the data forward process, while the purple arrows indicate our modifications. Compared to the vanilla SA, we introduce the Grouped-query Q-Projector to extract diverse semantic information, providing a favorable trade-off between performance and efficiency.

Table 2: Grasp detection performance comparison under dense and sparse annotation scenarios. 'MAB with SSA' and 'MAB with GSA' denote MAB module is built upon **S**ingle-query **S**hunted **A**ttention (SSA) and **G**rouped-query **S**hunted **A**ttention (GSA) respectively. The top two lines are trained *from scratch*. The GT-PCL denotes the two-stage training framework, which further adopts Pixel-wise Contrastive Learning (PCL) for the encoder. Empirical results demonstrate that the GSA consistently outperforms SSA proposed by [5] in both settings.

Method	Input Modality	Cornell		Jacquard	
		dense	sparse	dense	sparse
MAB with SSA	RGB-D	92.4	77.9	89.7	72.1
MAB with GSA	RGB-D	96.7	80.6	91.6	82.8
MAB with GSA + PCL (GT-PCL)	RGB-D	**99.2**	**89.3**	**97.3**	**84.3**

two-stage training framework GT-PCL, demonstrating the effectiveness of the encoder pre-train via Pixel-wise Contrastive Learning (PCL).

3 Detailed Ablation in MAB

To further explore the effectiveness of each layer in MAB, we split the discussion to separately address the impacts of the Shunted Attention Layer and Cross-token FFL. Specifically, we replace the Shunted Attention Layer and the Feed-forward Layer with the standard modules from the basic vision transformer as the basic baseline. Ablation results are summarised in the Table 3. Compared to traditional attention layers, the Shunted Attention Layer achieves a 12.1% performance improvement in the Jacquard [8] benchmark (67.4% vs 79.5%). Besides, empirical results indicate that the Cross-token FFL can significantly enhance model performance under the sparse annotation setting.

Table 3: Component ablation for Multi-scale Attention Blocks (MAB). We replace all components with the corresponding layers in the conventional ViT [9] as the baseline to validate the effectiveness of each module.

Shunted Attention Layer	Cross-token FFL	Cornell [7]		Jacquard [8]
		IW	OW	
		72.3	70.9	67.4
✓		81.6	77.8	79.5
	✓	74.3	71.1	72.6
✓	✓	**89.3**	**85.6**	**84.3**

Table 4: Hyper-parameter ablation for the distance threshold η in the PCL pre-training.

Distance Threshold η	Cornell [7]		Jacquard [8]
	IW	OW	
$\eta = 0.2$	79.5	78.2	73.3
$\eta = 0.3$	85.7	84.2	81.4
$\eta = 0.4$	**89.3**	**85.6**	**84.3**
$\eta = 0.5$	86.2	80.1	77.6

Compared with various attention architectures, the Shunted Attention Layer can effectively alleviate performance degradation caused by the sparse annotation.

4 Hyper-parameter Ablation in PCL

The setting of the distance threshold η in PCL is related to the resolution of the input raw image and the feature map. Specifically, during the preprocessing stage, data augmentation techniques such as center cropping and random shifting are applied to bring the input image resolution to 224x224. The feature map selected for Pixel-wise Contrastive Learning is resized to 28x28, resulting in a scaling ratio of 0.125. To ensure semantic consistency of pixels beyond the central pixel, the distance coefficient must be set within a reasonable range, no less than 0.125. Hence, the choice of distance threshold $\eta = 0.4$ is empirically determined. Table 4 examines the sensitivity to hyper-parameters of η.

For the ablation of distance threshold η, we set all related hyper-parameters to the following default values: input image resolution of 224×224, feature map of 28×28 and $\eta \in \{0.2, 0.3, 0.4, 0.5\}$.

5 Ablation study for Training Strategy

The PCL stage serves as a pre-training process in our framework. As depicted in the overview of our proposed GT-PCL, during the second stage, we freeze the encoder's parameters and solely update the parameters of the decoder, which is mentioned in the last sentence of Section C. This strategy aligns with our sparse annotation setting to optimize the grasp transformer with a few supervised labels. The reasons why we choose such strategy can be summarised into three manifolds:

- **Improved Generalization Performance.** As suggested by a line of studies in

Table 5: Ablation study for Training Strategy. There are three training strategies: (i) one-stage supervised training without PCL pre-train (GT); (ii) the encoder initially training through the PCL pre-train, during the second stage, we **jointly optimize both the encoder and decoder** of the grasp transformers (GT-PCL [J]); (iii) the encoder initially training through the PCL pre-train, during the second stage, we solely update the decoder parameters to avoid model overfitting (GT-PCL), which yeilding superior performance.

	Input Modality	Grasp Detection Accuracy (%)		
		GT	GT-PCL [J]	GT-PCL
Cornell [7]	RGB	82.5	83.7	**87.2**
	D	83.2	81.6	**85.8**
	RGB-D	83.5	85.4	**89.3**
Jacquard [8]	RGB	73.6	77.1	**80.5**
	D	71.8	79.4	**82.4**
	RGB-D	75.3	78.4	**81.6**

robotic manipulation [10, 11, 12], freezing the encoder layers often yields better generalization performance in tasks involving complex environments with unseen objects. As mentioned in the [12], freezing these layers can effectively prevent the overfitting of the encoder from the noise present in the training data, which is particularly beneficial when working with sparse annotations. Furthermore, it can alleviate so-called gradient conflicts, which means the gradient directions point away from each other, *e.g.*, appears a negative cosine similarity between the optimization objectives of self-supervised representation learning and supervised grasp regression, which severely degrade model performance in scenarios with limited annotations.

- **Computational Efficiency.** Freezing the encoder significantly reduces the number of parameters that need updating. Since the decoder only consists of stacked transported convolutional layers and activation functions, which require fewer parameters relative to the attention-based encoder, this setup not only simplifies the training process but also accelerates the model convergence to better performance with sparse annotations.

- **Empirical Validation.** We empirically validated this approach across two public benchmarks. Experimental results illustrate that models with a frozen pre-trained encoder significantly outperform the encoder jointly training with the decoder, which shows the effectiveness of our two-stage training strategy. Specifically, we compare our strategy with two widely adopted training schemes: (i) one-stage training process without the encoder pertaining, which is the **GT** setting; (ii) we use the pre-train the encoder with the same PCL solution but jointly update the whole model, including the parameters of decoder and encoder, which is the **GT-PCL[J]** setting. As shown in Table 5, our strategy significantly outperforms the other solutions with a considerable margin in both benchmarks.

6 Case Study in Physical Robot

We conduct more comprehensive real-robot experiments to elucidate the impact of sparse annotations on model performance. In detail, we train the model on the Cornell dataset

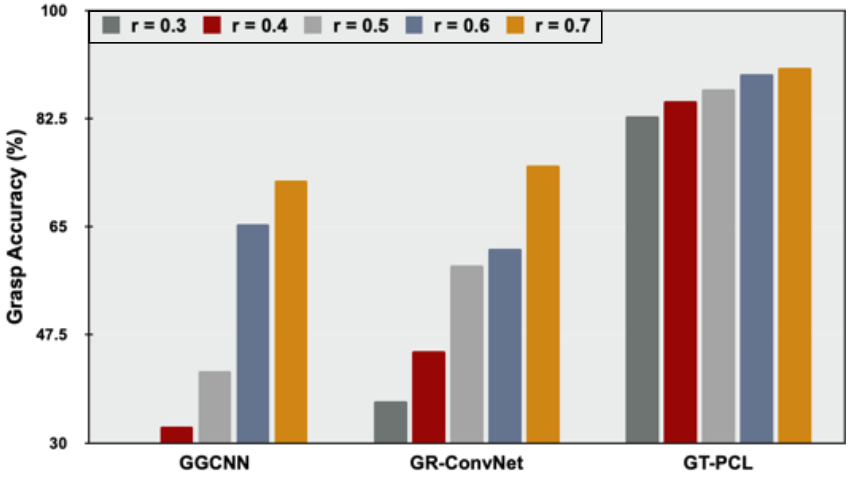

Figure 2: Performance evaluation under different annotation densities. We compare our GT-PCL with GGCNN [2] and GRConv-Net [4] to demonstrate the effect of removing different proportions of annotations on model performance, where r is the proportion of annotations used for model training.

with varying annotation densities and test it in physical scenarios. GGCNN [2] and GR-ConvNet [4] serve as baselines for comparative analysis. The experimental results, as illustrated in Figure 2, demonstrate that when only 30% of the annotations are removed, existing methods suffer a significant drop in performance. Conversely, our method maintains a grasp success rate of over 82% even with up to 70% of the annotations removed, significantly outperforming previous methods. This is attributed to our proposed PCL-based encoder pre-training and the introduction of the MAB module.

7 Experiments in Large-scale Grasp Dataset

Following the reviewer's suggestion, we provide additional experimental results on the large-scale REGRAD [13] benchmark. Specifically, REGRAD is a benchmark for relational grasping in dense clutter, which is built upon the well-known ShapeNet [14] dataset and contains 111.8k different scenes for model training and validation. Notably, the validation set can be split into two parts: the seen and unseen parts. For the unseen categories, the target objects are completely unknown in the training stage. This setting aims to validate the generalization performance of trained models, which is much more challenging. To verify the effectiveness of REGRAD [13], we select a part of data from 500 scenes to construct a validation set from the vanilla datasets as follows:

- **REGRAD-seen-val-500**: the first 500 scenes of REGRAD unseen-val set, consisting of 4500 images including different object instances of known categories.
- **REGRAD-unseen-val-500**: the first 500 scenes of REGRAD unseen-val set, consisting of 4500 images of unknown objects.

Furthermore, we use the same annotation sparsity method to construct the training datasets with sparse and dense annotations and evaluate the grasp performance under diverse numbers of annotations. Briefly, we randomly remove redundant grasp annotations and only retain 30% of the labels for model training under sparse annotation detection.

Table 6: Performance evaluation in the large-scale REGRAD [13] benchmarks. Under both sparse annotation and dense annotation settings, our GT-PCL achieves state-of-the-art performance with a significant advantage.

Method	Input	REGRAD-seen-val-500		REGRAD-unseen-val-500		Latency
		Dense (%)	Sparse (%)	Dense (%)	Sparse (%)	
GGCNN	RGB-D	53.3	32.8	47.6	21.4	19
GR-ConvNet	RGB-D	56.4	38.4	50.2	29.7	20
ORANGE	D	64.6	41.6	55.8	33.4	29
GP-Net	RGB	78.2	54.6	61.4	39.1	22
	RGB-D	82.6	59.2	73.2	48.6	23
GT-PCL	RGB	80.4	74.9	77.8	68.4	22
	RGB-D	**84.6**	**78.2**	**81.4**	**72.5**	23

Following the methods described in [13], we conduct experiments with diverse modal inputs, including RGB and depth images. We apply standard augmentation techniques for RGB images, identical to those used in [4] before feeding the images into the neural network. The rectangle metric [7] is used to report the performance of different methods. For a fair evaluation, we re-train GG-CNN [2], GR-ConvNet [4], ORANGE [15], and GP-Net [16] networks on the REGRAD training set as strong baselines. The results are summarized in the Table 6. Our method obtains a success rate of 84.6% and exhibits 81.4 generalization performance on the unseen validation. Compared with GG-CNN, our method achieves the grasp accuracy from 32.8% to 78.2%.

For the sparse annotations setting, our GT-PCL with a single RGB input can obtain better performance than these competitive multi-modal grasp detectors, such as GR-ConvNet and GP-Net [16]. Empirical results demonstrate that our GT-PCL achieves state-of-the-art grasp performance on the REGRAD dataset, significantly outperforming existing strong baselines. In addition, we also assess network performance across different input modalities. Table 6 bottom illustrates that our method performs better on multi-modal data than on single-modal data as the diverse input modalities can provide more useful information for representation learning. We also evaluate the time consumption of GT-PCL and other methods, reporting that GP-Net achieves the most advanced detection results on the REGRAD dataset with only a minimal additional latency of 1–2 ms. This indicates that our method is well-suited for real-time robotic grasping systems, which can provide faster response and smoother operation.

8 Ablation for PCL with Visualization Analysis

To validate the benefits of PCL, we visualized the grasp confidence maps and angle maps generated by our model. Figure 3 clearly depicts that model with PCL maintains high confidence in the graspable objects, effectively ignoring the disruptive background. On the contrary, models without the encoder pretraining tend to prioritize the textural information and assign high grasp confidence to disturbed background regions, which would mislead the grasp detectors. The ablation demonstrates that PCL encourages the model to focus on geometric invariance, thereby reducing the impact of irrelevant textural

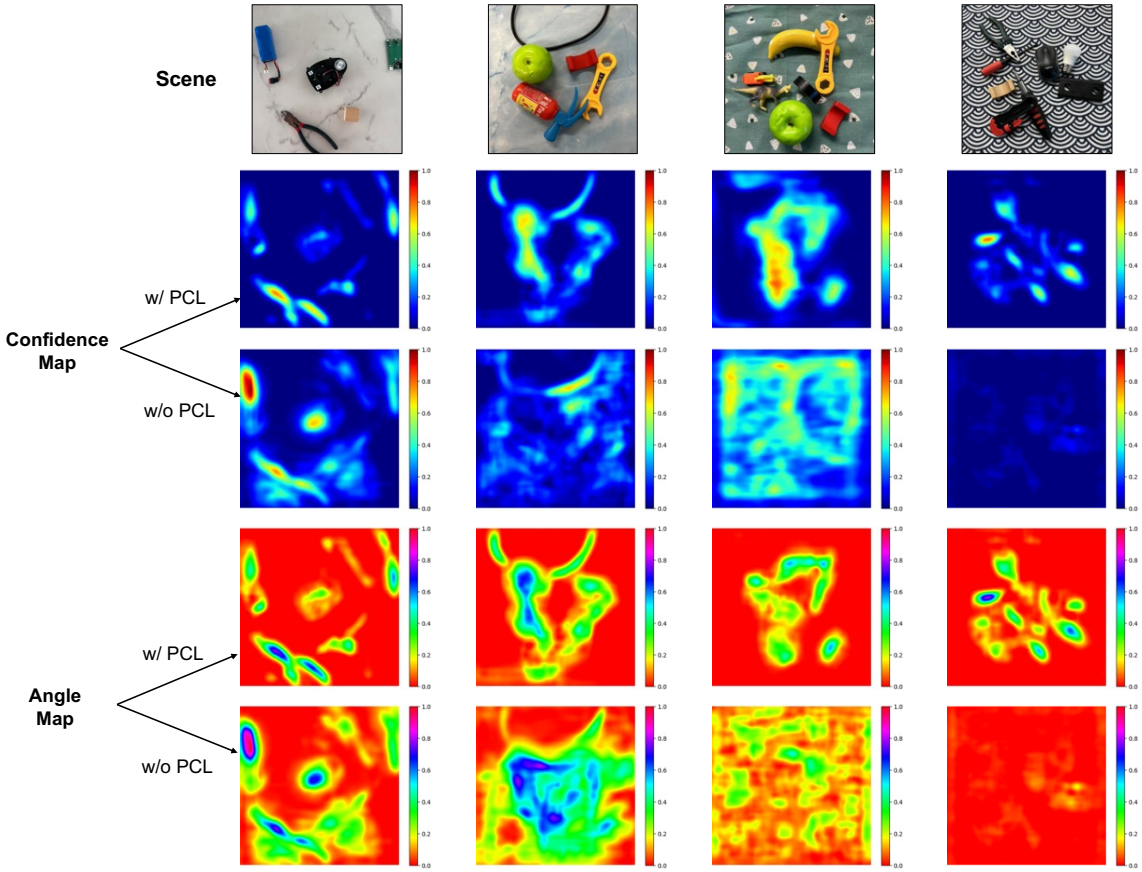

Figure 3: Visualization analysis on various physical platforms. we visualize the normalized confidence and angle maps obtained by the model with and without using PCL. The first row denotes using PCL, while the second row denotes results not using PCL.

information that can lead to overfitting or misinterpretation in complex environments. It is worth mentioning that the last two scenes were not encountered during the training stage. This indicates that the pre-training encoder based on PCL can effectively enhance the model's generalization capabilities, which is crucial for practical deployment in real-world settings.

9 Conclusion

The supplementary experimental results and visualizations provided in this file demonstrate the robustness and effectiveness of our proposed method under various experimental conditions. We believe these additions significantly strengthen the contributions of our work. Thank you for considering our revised manuscript. We look forward to your constructive comments and feedback.

References

[1] U. Asif, J. Tang, and S. Harrer, "Graspnet: An efficient convolutional neural network for real-time grasp detection for low-powered devices," in *IJCAI*, vol. 7, pp. 4875–4882, 2018.

[2] D. Morrison, P. Corke, and J. Leitner, "Learning robust, real-time, reactive robotic grasping," *IJRR*, vol. 39, no. 2-3, pp. 183–201, 2020.

[3] R. Xu, F.-J. Chu, and P. A. Vela, "Gknet: grasp keypoint network for grasp candidates detection," *The International Journal of Robotics Research*, vol. 41, no. 4, pp. 361–389, 2023.

[4] S. Kumra, S. Joshi, and F. Sahin, "Antipodal robotic grasping using generative residual convolutional neural network," in *IROS*, pp. 9626–9633, IEEE, 2020.

[5] S. Ren, D. Zhou, S. He, J. Feng, and X. Wang, "Shunted self-attention via multi-scale token aggregation," in *CVPR*, pp. 10853–10862, 2022.

[6] J. Ainslie, J. Lee-Thorp, M. de Jong, Y. Zemlyanskiy, F. Lebrón, and S. Sanghai, "Gqa: Training generalized multi-query transformer models from multi-head checkpoints," *arXiv preprint arXiv:2305.13245*, 2023.

[7] Y. Jiang and S. Moseson, "Efficient grasping from rgbd images: Learning using a new rectangle representation," in *ICRA*, pp. 3304–3311, 2011.

[8] A. Depierre, E. Dellandr'ea, and L. Chen, "Jacquard: A large scale dataset for robotic grasp detection," in *IROS*, pp. 3511–3516, 2018.

[9] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, M. Unterthiner, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *ICLR*, 2020.

[10] S. Karamcheti, S. Nair, A. S. Chen, T. Kollar, C. Finn, D. Sadigh, and P. Liang, "Language-driven representation learning for robotics," *arXiv preprint arXiv:2302.12766*, 2023.

[11] S. Chen, R. Garcia, I. Laptev, and C. Schmid, "Sugar: Pre-training 3d visual representations for robotics," *arXiv preprint arXiv:2404.01491*, 2024.

[12] I. Radosavovic, T. Xiao, S. James, P. Abbeel, J. Malik, and T. Darrell, "Real-world robot learning with masked visual pre-training," in *Conference on Robot Learning*, pp. 416–426, PMLR, 2023.

[13] H. Zhang, D. Yang, H. Wang, B. Zhao, X. Lan, J. Ding, and N. Zheng, "Regrad: A large-scale relational grasp dataset for safe and object-specific robotic grasping in clutter," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 2929–2936, 2022.

[14] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, *et al.*, "Shapenet: An information-rich 3d model repository," *arXiv preprint arXiv:1512.03012*, 2015.

[15] G. Chalvatzaki, N. Gkanatsios, P. Maragos, and J. Peters, "Orientation attentive robotic grasp synthesis with augmented grasp map representation," *arXiv preprint arXiv:2006.05123*, 2020.

[16] Y. Yang, Y. Xing, J. Zhang, D. Tao, *et al.*, "Gp-net: A lightweight generative convolutional neural network with grasp priority," *APSIPA Transactions on Signal and Information Processing*, vol. 12, no. 1, 2023.