2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

Tool-Grasp: A 6-DoF Functional Grasping Framework for General-Purpose Hand Tools

Hongliang Lei[1], Jian Huang[1], Andong Li[1], Haoyuan Wang[1], Chen Liu[1], Wei Luo[2], and Jiuyao Xiang[3]

Abstract— **Detecting functional grasp poses for tool operation is critical for robots in complex real-world tasks, yet existing methods lack this capability. Key challenges are: 1) Scarce real-world datasets with fine-grained functional labels and task-valid grasp annotations, as their construction requires domain knowledge (making annotation labor-intensive/subjective) and linking poses to tool usage (beyond stability checks); 2) Difficulty in fine-grained functional segmentation, where minimal sub-region differences are overwhelmed by global cues/noise, with 3D model-dependent methods impractical in unstructured settings; 3) Poor 6-DoF grasp alignment with functional regions due to high morphological heterogeneity, as existing methods either fail to balance stability and functional constraints (high-score grasps outside regions) or are limited to low degrees of freedom. To address these, we build the Tool-Grasp Dataset (20 tool categories, 50 scenes, 12,600 RGB-D images, 250M+ 6-DoF annotations) with fine-grained functional labels. We propose Tool-Grasp, a two-stage 6-DoF framework: Stage 1's Mask-Guided Grasp Region Segmentation Network (MG-GRSN) leverages tool-specific semantics to output precise functional masks, mitigating intra-tool variability; Stage 2's Quality-Aware Multi-Modal Grasp Pose Detection Network (QAM-GPDN) uses these masks to constrain predictions, fusing RGB-D features with a quality module to select aligned poses. Experiments show MG-GRSN outperforms baselines by 3.5% (seen) and 5.2% (unseen) in mIoU; QAM-GPDN boosts functional pose AP by 2.89% (seen) and 3.76% (unseen). Real-robot experiments validate real-world effectiveness.**

I. INTRODUCTION

Object grasping is a core robotic capability, underpinning diverse manipulation tasks in industrial, precision manufacturing, and household scenarios. Unlike generic objects—where stable clamping suffices—tool-like objects require prioritizing handle grasping (among multiple graspable regions) to align with their functional intent. Successful tool grasping demands both anti-slip grasp stability and precise grasp pose optimization (with focus on handles) to enable seamless integration into subsequent tasks [1]–[3]. Our work targets General-Purpose Hand Tools, which feature distinct handle-functional region structures (e.g., daily service tools: knives, forks, and spoons; light-duty work tools: hammers, screwdrivers, and wrenches) and excludes small precision tools (primarily relying on fingertip control).

High-quality annotation data is foundational for functional grasping of tool-like objects, enabling advances in related algorithms [4]. Existing datasets provide valuable resources: NYU Depth V2 [5] and SUN RGB-D [6] offer RGB-D data with instance-level indoor labels for multi-modal fusion; ScanNet [7] adds 2D/3D joint annotations for scene reconstruction; GraspNet-1Billion [8] provides dense 6-DoF grasp annotations to bridge simulation-to-real gaps; and Sim-Grasp [9] uses physics simulation for validation. However, none include fine-grained tool functional region labels or task-valid grasps. Wang *et al.* [10] proposed a task-oriented 6-DoF dataset but it is in simulation, leading to real-world discrepancies. Building a real-world dataset faces key challenges: functional region annotation requires domain knowledge, with task-critical part distinction being labor-intensive and subjective; task-valid grasp annotation demands linking poses to tool usage, adding complexity beyond generic stability checks. Thus, a real-world dataset with fine-grained functional region and 6-DoF grasp annotations is urgently needed.

Accurate part-level segmentation of tool functional regions is a critical bottleneck, as it supplies priors for grasp pose constraint and adaptive force control [11]. While 3D part segmentation methods attain high accuracy, they rely on precomputed 3D models or multi-view point clouds—impractical in unstructured real-world settings. Recent advances (SGNet [15] with space-guided convolutions; DBCAN [16] for boundary ambiguity; HDBFormer [17] balancing local/global features; DFormerv2 [18] using geometric self-attention) refine spatial adaptability, but a core challenge remains: fine-grained discrimination. Tool functional sub-regions often have minimal structural differences, which are easily overwhelmed by global semantic cues. Additionally, real-world noise further blurs these subtle differences, leading to misclassification of non-functional regions as functional. Thus, developing segmentation algorithms that capture such fine-grained features remains urgent.

Accurate alignment between 6-DoF grasp poses and tool functional regions, which is defined by two core conditions, remains a critical bottleneck even with reliable segmentation

*This work was supported in part by Hubei Science and Technology Major Project under Grant 2024BAA007, and in part by the National Natural Science Foundation of China, Ye Qisun Science Foundation: U2341228. (Corresponding authors: Jian Huang; Wei Luo.)

[1]Hongliang Lei, Jian Huang, Andong Li, Haoyuan Wang, and Chen Liu are with Hubei Key Laboratory of Brain-Inspired Intelligent Systems, School of Artificial Intelligence and Automation, Huazhong University of Science and Technology (HUST), Wuhan, Hubei 430074, China, and also with the Key Laboratory of Image Processing and Intelligent Control, School of Artificial Intelligence and Automation, Huazhong University of Science and Technology (HUST), Wuhan, Hubei 430074, China leihl@hust.edu.cn; huang_jan@mail.hust.edu.cn; li_and@hust.edu.cn; why427@hust.edu.cn; 2621513841@qq.com

[2]Wei Luo is with the Department of Innovation Center, China Ship Development and Design Center, Wuhan, Hubei 430064, China csddc_weiluo@163.com

[3]Jiuyao Xiang is with the Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology, Hong Kong, 999077, China jxiangag@connect.ust.hk

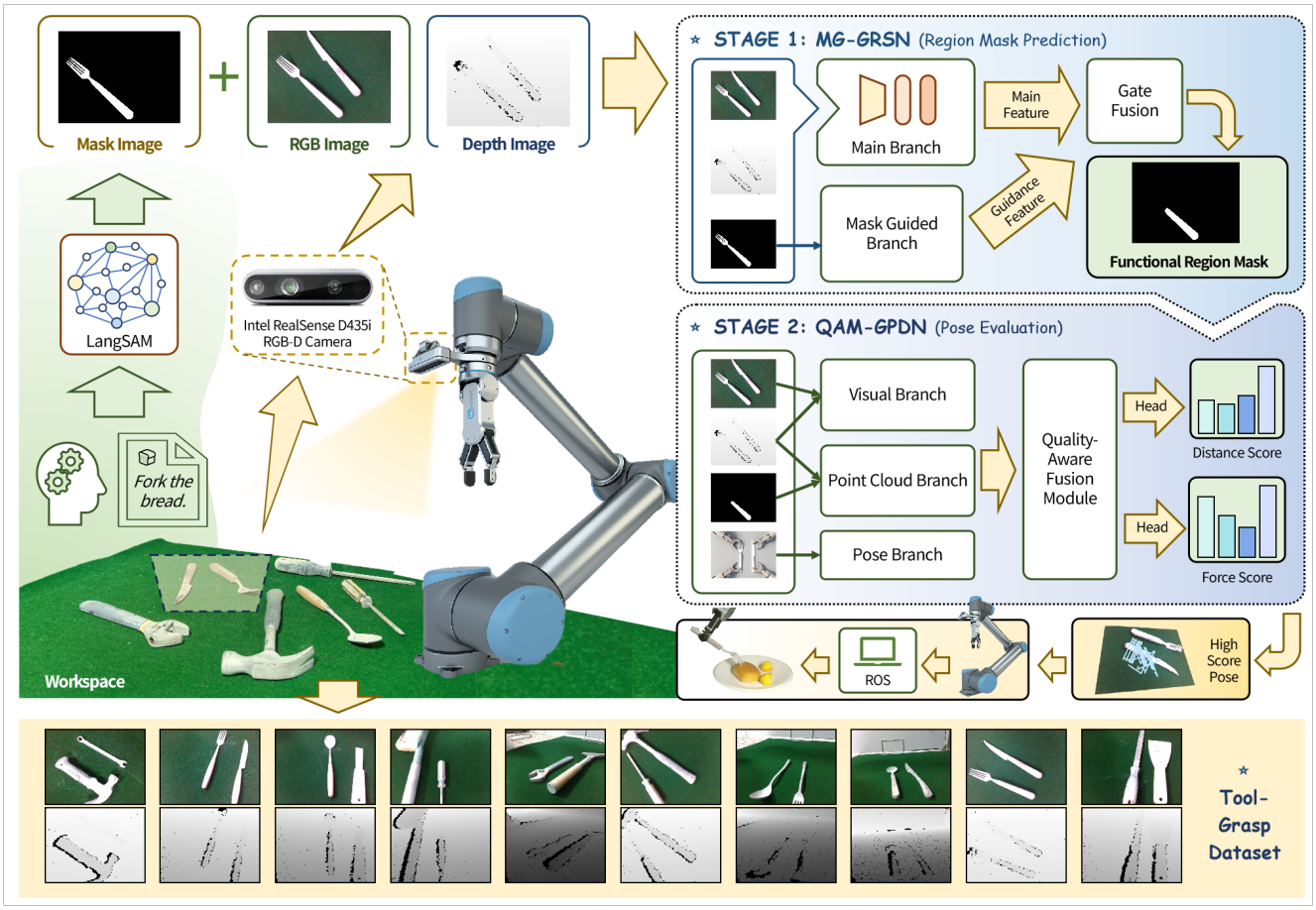


Fig. 1. The overall framework of Tool-Grasp. After selecting the target tool, the LangSAM [14] model outputs the tool's global mask. In stage 1, this mask—along with RGB and depth images—is fed into the Mask-Guided Grasp Region Segmentation Network (MG-GRSN). The network generates guidance features to refine the main features with the output being a pixel-level functional region mask. In stage 2, the Quality-Aware Multi-Modal Grasp Pose Detection Network (QAM-GPDN) extracts visual, point cloud-based, and pose-based features. These features are then fused in the Quality-Aware Fusion Module, after which two hierarchical uncertainty heads predict force scores and distance scores. Finally, the robot grasps the target tool using a high-score pose and completes subsequent operations.

of tool functional regions [19]. These two conditions are: 1) the grasp pose itself is mechanically stable, and 2) all contact points of the grasp lie within the tool's functional regions. A common approach involves generating global 6-DoF grasps and then filtering them via functional region data. Recent works focus on multi-modal fusion: Gou *et al.* [22] boosted robustness to low-quality depth; GraNet [23] used structure-aware attention for 6-DoF grasp detection; HGGD [24] generated diverse poses via global-to-local guidance. However, a key limitation persists: high-scoring grasps often have contact points outside functional regions. While GraspCLIP [12] generates functional region-aligned grasps, its outputs are limited to low degrees of freedom (low-DoF), which constrains stability in dynamic tasks. The core challenge centers on geometric diversity: Tool functional regions exhibit extreme morphological heterogeneity, and existing methods rely on generic alignment criteria that fail to balance stability and functional constraints. For instance, a criterion prioritizing stability may force grasps outside functional regions, while a criterion focusing on functional regions may compromise stability. This mismatch blocks reliable 6-DoF functional grasp alignment. Thus, developing methods for 6-DoF grasps precisely aligned with functional regions in few-object scenes remains our primary focus.

The main contributions of this work are as follows:

- **Tool-Grasp Dataset**: We present a real-world dataset featuring 20 tool categories, 50 scenes, 12,600 RGB-D images, and over 250 million 6-DoF annotations. It is designed with fine-grained annotations for grasping to address the scarcity of data on tool functional grasping.
- **MG-GRSN**: We propose a functional region segmentation network centered around dynamic mask guidance. This network leverages global object masks and refines them through multi-scale processing, enabling simultaneous capture of fine local edge details and global contour structures. A cross-attention mechanism facilitates the fusion of multi-scale mask features, while a spatial attention module generates dynamic spatial weights. This dynamic mask-guided framework substantially improves the accuracy of pixel-level functional region

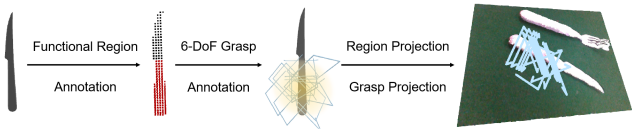


Fig. 2. The processes of functional region annotation and 6-DoF grasp annotation. First, each tool's 3D mesh is segmented into functional regions, which are then projected onto corresponding RGB-D images. Next, grasp poses are sampled such that their centers align to functional regions on the 3D mesh; these poses are then validated through physics-based simulation to ensure structural feasibility. Finally, the validated poses are projected onto RGB-D images to complete the 6-DoF grasp annotation process.

segmentation.

- **QAM-GPDN**: We propose a region-aware pose detection network that leverages functional region masks to filter raw point clouds, effectively mitigating background point interference and facilitating targeted extraction of 3D geometric features for functional grasp regions. Furthermore, QAM-GPDN incorporates a dynamic weight fusion mechanism: it first predicts quality metrics for each modality, and then scales the base attention weights accordingly using these metrics to dynamically adjust modal contributions. This design enables QAM-GPDN to output 6-DoF grasps that are precisely aligned with functional regions.

Fig. 1 shows the framework of our proposed methods. The rest of this article is organized as follows: Section II introduces the constructed Tool-Grasp Dataset the constructed dataset and details the data collection and annotation processes. Section III describes the proposed methods in detail. Section IV presents the experimental setup, results, and analysis. Finally, Section V concludes this article.

II. TOOL-GRASP DATASET

We propose a dataset—named Tool-Grasp—with dense and rich annotations for functional region segmentation grasp pose prediction named Tool-Grasp. Our dataset contains 20 common tool categories covering daily-use functional tools. For tools with intra-category variations, multiple variants are included to capture morphological diversity. Data were collected from 50 scenes, with 252 RGB-D images per scene (12,600 images in total) captured from diverse viewpoints. Each image is paired with global object masks, functional region masks, and 6-DoF object poses. Over 250 million 6-DoF grasp poses are annotated, each of which is labeled with force closure scores and geometric metrics. Our dataset further provides high-fidelity 3D mesh models, enabling cross-modal alignment between 2D images and 3D geometries for both annotation and evaluation.

A. Data Collection

Inspired by [8], we automate data collection to ensure consistency while capturing real-world variability. A camera is mounted on the robotic arm, and both the camera and arm are calibrated to obtain the homogeneous transformation matrix between the camera frame and the end-effector frame ($P_{c\text{-}e}$). The robotic arm moves along a predefined trajectory covering a quarter-sphere around the workspace, capturing images from 252 unique viewpoints per scene. For each viewpoint, the end-effector pose ($P_{e\text{-}b}$) relative to the arm base frame is recorded, enabling cross-viewpoint pose consistency. Each scene contains two randomly selected tools placed on a workspace table. Additional details are provided in the video included in the supplementary materials.

B. Data Annotation

Annotations are generated through a hybrid pipeline combining 6D object pose annotation, functional region annotation, and 6-DoF grasp annotation, ensuring accuracy and functional relevance. The processes of functional region annotation and 6-DoF grasp annotation are shown in Fig. 2.

1) 6D Object Pose Annotation: To obtain the 6D pose of tools in each image ($P_{o\text{-}c}$), we select one representative image per scene (with distinct tool features). We establish manual correspondences between 2D image points and 3D mesh vertices, and use the Perspective-n-Point (PnP) algorithm [13] to initialize $P_{o\text{-}c}^{init}$. To obtain $P_{o\text{-}c}^{final}$ for all other images in the scene, we propagate the refined pose $P_{o\text{-}c}^{init}$ using the recorded $P_{e\text{-}b}^{init}$ and $P_{e\text{-}b}^{final}$:

$$P_{o\text{-}b} = P_{o\text{-}c}^{init} \cdot P_{c\text{-}e} \cdot P_{e\text{-}b}^{init} \tag{1}$$

$$P_{o\text{-}c}^{final} = P_{o\text{-}b} \cdot \left(P_{e\text{-}b}^{final}\right)^{-1} \cdot (P_{c\text{-}e})^{-1} \tag{2}$$

where $P_{o\text{-}b}$ is the object pose in the arm base frame (constant per scene).

2) Functional Region Annotation: Functional regions are predefined according to the tool's functionality. For each tool, its 3D mesh is first manually segmented,after which only the functional regions are retained. Using the 6D object pose $P_{o\text{-}c}$, the 3D functional region is projected onto the RGB-D image plane to generate pixel-level functional masks. Global object masks are derived by projecting the full 3D mesh of the tool, which enables effective background suppression.

3) 6-DoF Grasp Annotation: Grasp poses are generated in 3D space and projected onto 2D images using $P_{o\text{-}c}$, ensuring alignment with functional regions. Uniform surface points are sampled exclusively on the functional regions of the tool's 3D mesh model. For each point, grasp translation vectors are sampled on concentric spheres around the point, and rotations are generated such that the gripper opening faces the point. Additional rotations around the gripper's central axis are added to capture orientation variations. Grasps are validated in simulation: those colliding with non-functional regions are discarded, and remaining poses are scored using force closure metrics. Finally, the remaining poses are projected into the scene. As proposed in [8], the force closure score is defined as:

$$s = 1.1 - \mu_{\min} \tag{3}$$

where $\mu_{\min}$ is the minimum friction coefficient required for stable grasping. Valid 3D grasps are projected to each image using $P_{o\text{-}c}$, resulting in image-specific 6-DoF grasp labels:

$$T^i = P_{o\text{-}c}^i \cdot T \tag{4}$$

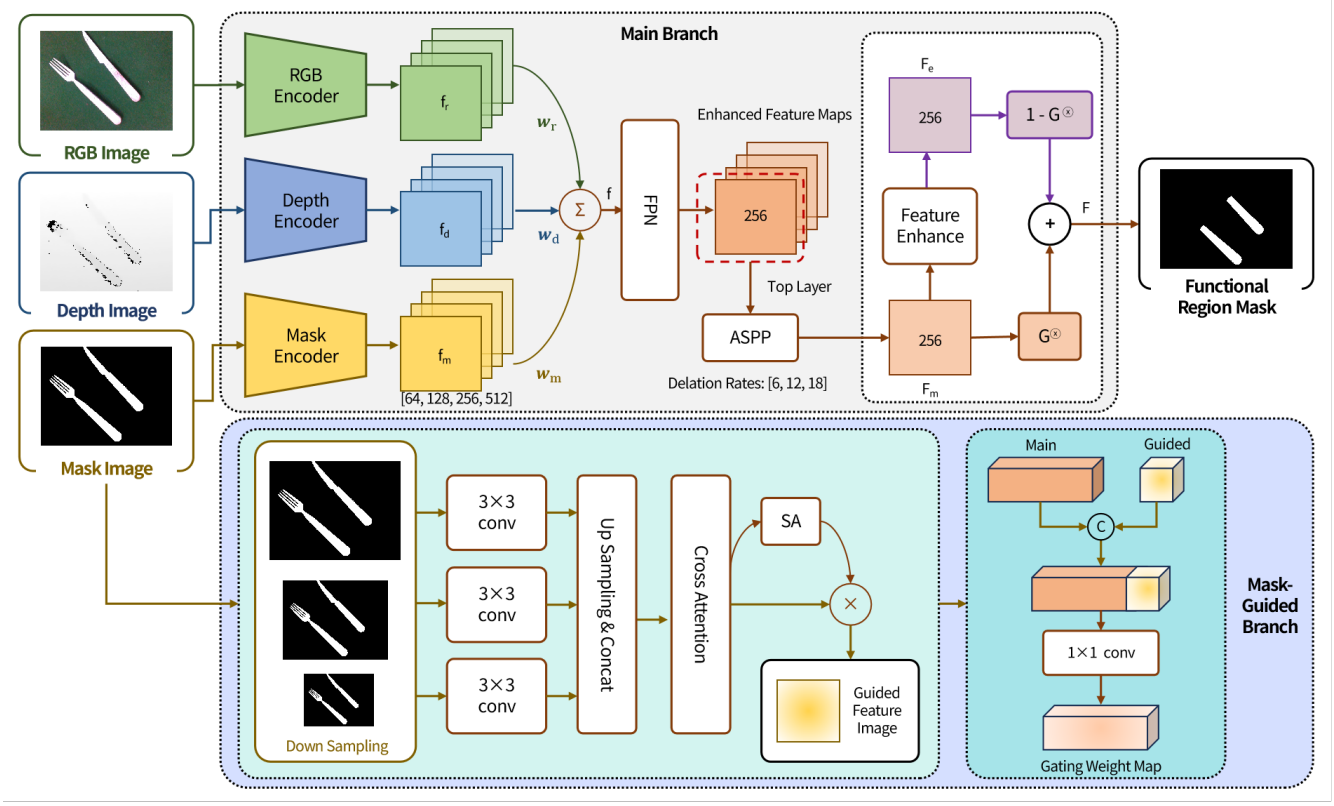


Fig. 3. Overview of our MG-GRSN. The network takes an RGB image, depth image, and global object mask as inputs. The Main Branch employs three lightweight ResNet-18 encoders with stage-wise dynamic weighting to fuse multi-modal features, followed by FPN and ASPP modules for multi-scale context enhancement. The Mask-Guided Branch processes the input global object mask at multiple scales, applies cross-attention to achieve scale-aware feature fusion, and generates a guided feature via spatial attention. Main features (from the Main Branch) and guided features are combined through a gating mechanism to predict the pixel-wise functional grasp region mask.

where T is the 3D grasp pose in the tool's object frame. Each projected grasp is paired with its force closure score s and the average distance d between the contact points and the tool's centroid.

III. METHODOLOGY

The Tool-Grasp framework consists of two networks: the MG-GRSN generates guidance features to refine main features, outputting pixel-level functional region masks; the QAM-GPDN extracts visual, point cloud, and pose features, fuses them, and predicts force scores, distance scores, and their uncertainty estimates. Below, we elaborate on these two networks in detail.

A. Mask-Guided Grasp Region Segmentation Network

MG-GRSN aims to predict pixel-level masks of functional grasp regions by leveraging global object masks to guide multi-modal feature learning. The network structure is shown in Fig. 3.

1) Main Branch: To satisfy real-time deployment requirements for robotic systems, we employ three lightweight ResNet-18 encoders [21] (with fully connected layers removed, retaining only the first four stages) to extract hierarchical features. These features are denoted as f_r, f_d, and f_m, and they correspond to RGB, depth, and global object mask inputs respectively. Each encoder outputs four feature maps characterized by increasing semantic abstraction, decreasing spatial resolution, and fixed channel dimensions of [64, 128, 256, 512]. To address dynamic variations in modal reliability, we propose a stage-specific dynamic weight network. For each feature stage, these three modal features are first concatenated and then processed through a 1×1 convolution layer to generate normalized weights (with normalization enforced by the Softmax function) for the RGB, depth, and mask modalities. The fused feature at each stage is computed as follows:

$$f = w_r \cdot f_r + w_d \cdot f_d + w_m \cdot f_m \tag{5}$$

where w_r, w_d, and w_m represent adaptively learned weights that adjust to reliability of their respective modalities.

Functional grasp regions exhibit significant scale variations. To address this, we employ a Feature Pyramid Network (FPN) to fuse multi-scale features: lateral convolutions first unify the 4-stage fused features to 256 channels, followed by top-down upsampling to merge high-resolution details (from shallow layers) with semantic information (from deep layers), generating 4 enhanced feature maps at different scales. To capture global context, an Atrous Spatial Pyramid Pooling (ASPP) [25] module is applied to the top-layer pyramid feature. The ASPP uses 3×3 convolutions with dilation

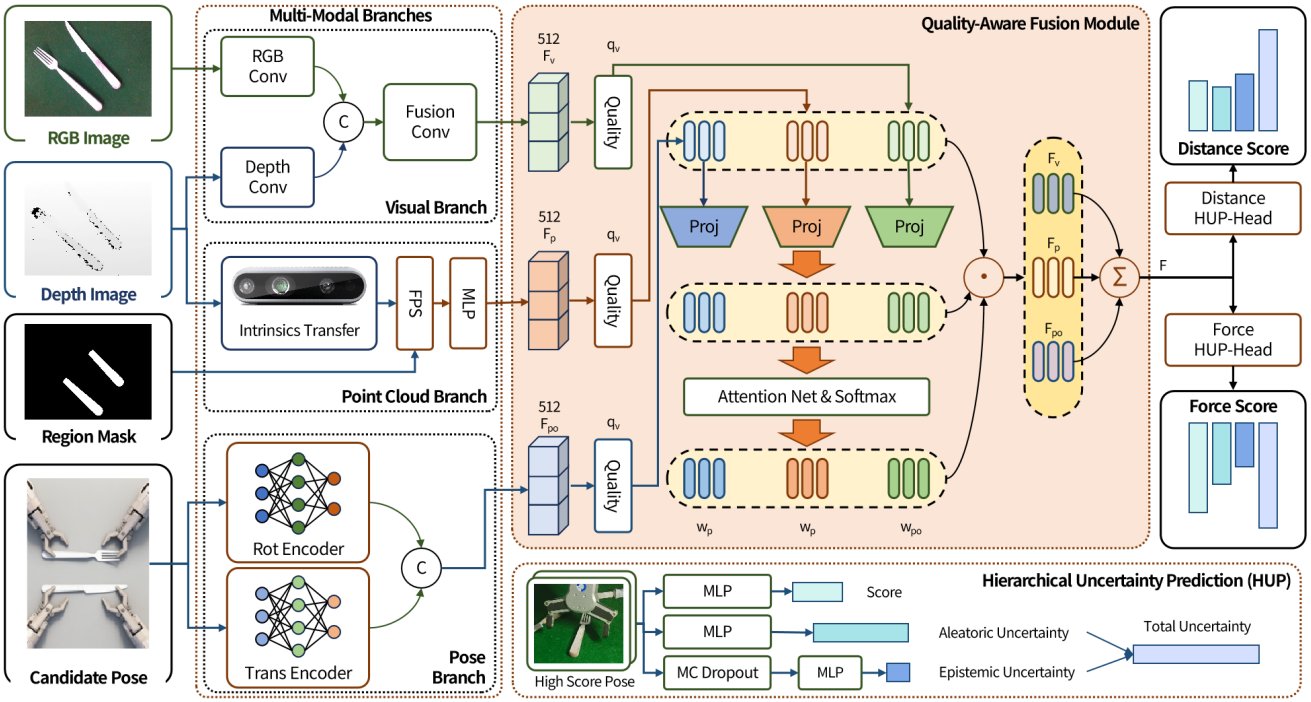


Fig. 4. Architecture of the proposed QAM-GPDN. The network takes an RGB image, depth image, functional region mask, and candidate 6-DoF grasp poses as inputs. Three branches extract visual, point cloud, and pose features. A quality-aware fusion module combines them using quality-weighted attention, followed by hierarchical uncertainty prediction to output distance scores, force scores, and uncertainty estimates.

rates of [6, 12, 18] to extract multi-scale contextual features, while integrating global average pooling to capture scene-level information. This joint operation yields a 256-channel aggregated feature that effectively enhances the model's robustness. Ultimately, this process produces the main feature F_m for subsequent functional region segmentation tasks.

2) Mask-Guided Branch: This branch directs learning toward the target object region to avoid background interference. To address edge noise and scale bias in the global object mask, the input mask is first downsampled to multiple scales. Each scaled mask is processed using 3×3 and 5×5 convolutions to extract contour features, then upsampled back to the original size—thereby capturing both local edge details and global contour structures. To model long-range dependencies between mask features across scales, we introduce a cross-attention mechanism: multi-scale mask features are reshaped into a sequence format, and a multi-head self-attention layer (with 4 heads) is applied along the sequence dimension to fuse scale-specific features. This enables information exchange between fine and coarse scales, effectively enhancing the representation of ambiguous regions. After attention, the features are averaged across scales to produce a unified mask feature map. To dynamically weight spatial regions based on task relevance, we compute a spatial attention (SA) weight map by fusing the unified mask feature map with the raw mask (after smoothing via a 3×3 convolution), yielding the guided feature. This guided feature is then combined with F_m to generate a gating weight map. Using this gating weight map G, the final feature is obtained as:

$$F = F_m \cdot G + F_e \cdot (1 - G) \tag{6}$$

where F_e denotes the output of refining F_m via a convolution layer.

To train the MG-GRSN, we adopt binary cross-entropy (BCE) loss integrated with spatial masking. Specifically, we introduce a binary valid mask $M \in \{0,1\}^{H \times W}$, where $M_{i,j} = 1$ indicates that pixel (i, j) belongs to the tool region. The loss function is:

$$\mathcal{L} = -\frac{1}{\sum M} \sum_{i,j} \Big((t_{i,j} \log p_{i,j} + (1 - t_{i,j}) \log(1 - p_{i,j})) \cdot M_{i,j} \Big) \tag{7}$$

where $p_{i,j}$ represents the model's predicted probability of the grasp region at pixel (i, j), and $t_{i,j}$ denotes the ground-truth grasp mask. The term $\sum M$ (sum of all valid pixels in M) normalizes the loss by the number of tool-region pixels.

B. Quality-Aware Multi-Modal Grasp Pose Detection Network

QAM-GPDN evaluates candidate 6-DoF grasp poses, selecting those within functional regions and outputting: force score $S_f \in [0, 1]$ (stability), distance score $S_d \in \mathbb{R}$ (distance), and uncertainty estimates (for reliability assessment). The network structure is shown in Fig. 4.

1) Multi-Modal Branches: Three parallel branches extract complementary features: Visual branch captures texture and

geometry from RGB-D images. RGB and depth images are processed via 3×3 convolutions, ReLU activations, and max-pooling (stride=2) to generate 256-D fused features, which are refined via 3×3 convolutions and adaptive average pooling to output 512-D global visual features. Point cloud branch extracts 3D geometric details from functional regions. Using camera intrinsics, depth maps are converted to 3D point clouds in camera coordinates. 1024 points are sampled via farthest point sampling (FPS) from the functional mask region to preserve critical geometric structures. These points are processed with 1D convolutions and max-pooling to generate 512-D global point cloud features. Pose branch models spatial relationships of grasp poses. The 4×4 pose matrix is decomposed into rotation (3×3, flattened to 9D) and translation (3D) components, encoded via separate MLPs, and concatenated to form 512-D pose features. This separation enhances modeling of the independent effects of rotation (contact angle) and translation (relative position) on grasp quality.

2) Quality-Aware Fusion Module: To address uneven reliability across modalities, we introduce a quality-aware fusion mechanism: MLPs predict quality metrics for each modality: q_v (RGB-D SNR), q_p (point cloud density and noise), and q_{po} (spatial alignment with the functional mask). To enable cross-modal fusion, input features of varying dimensions are projected to a shared latent space. Attention networks compute base weights w_v, w_p, and w_{po} from feature correlations, which are scaled by quality scores to generate final weights:

$$F = q_v \cdot w_v \cdot F_v + q_p \cdot w_p \cdot F_p + q_{po} \cdot w_{po} \cdot F_{po} \quad (8)$$

where F_v, F_p, and F_{po} denote features derived from the visual, point cloud, and pose branches, respectively.

3) Hierarchical Uncertainty Prediction: This module explicitly decomposes uncertainty into aleatoric and epistemic components. Specifically, it estimates the score via one 2-layer multi-layer perceptron (MLP), and aleatoric uncertainty via another distinct 2-layer MLP. For epistemic uncertainty prediction, Monte Carlo (MC) Dropout is applied prior to the input layer of an additional MLP, enabling stochastic uncertainty quantification through forward-pass sampling. The total uncertainty σ_t is defined as the square root of the sum of the variances of the two components, which guides threshold-based decision-making (e.g., discarding poses with $\sigma_t > 0.3$).

QAM-GPDN is trained using a negative log-likelihood (NLL) loss with uncertainty regularization, formulated as:

$$\mathcal{L} = \frac{1}{2} \sum \left(\log\left(\sigma_t^2\right) + \frac{(y - \hat{y})^2}{\sigma_t^2} \right) + 0.01 \cdot \left(\log\left(\sigma_t^2\right) + \sigma_t^2 \right) \quad (9)$$

where $\hat{y}$ denotes the predicted score and y is the corresponding ground truth. This loss balances prediction accuracy and uncertainty calibration, with the 0.01-scaled regularization term preventing overconfident or undercalibrated uncertainty estimates.

IV. EXPERIMENTS

To validate the effectiveness of our proposed Tool-Grasp framework, we conducted comprehensive experiments including benchmark comparisons, ablation studies, and real-robot evaluations—specifically on a UR5e robot equipped with an Intel RealSense D435i depth camera and an OnRobot RG2 gripper.

A. Implementation Details

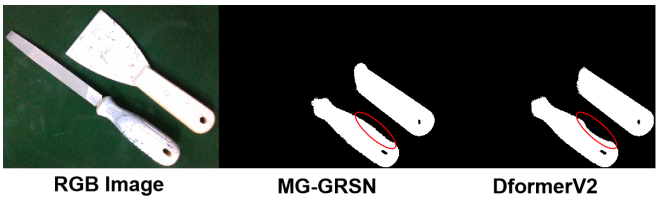
All experiments were conducted on the proposed Tool-Grasp Dataset or real robots. For benchmarking, we split the dataset into "seen" (16 tool categories) and "unseen" (4 novel tool categories) subsets to evaluate the framework's generalization capability. For functional region segmentation, we used mean Intersection over Union (mIoU) as the primary metric, which is calculated as the average IoU between predicted functional region masks and ground-truth annotations within the global mask of the target object; for 6-DoF grasp pose detection, following standard practice in 6D grasp evaluation [8], we adopted Average Precision (AP) as the evaluation metric. MG-GRSN was implemented in PyTorch and trained on a single NVIDIA RTX 3090 GPU, with mixed-precision training adopted to accelerate the training process. Data augmentation strategies included random scaling (0.9–1.1×), 90° in-plane rotations, horizontal/vertical flips, random shadow addition, and motion blur. The network was optimized using the AdamW optimizer, with an initial learning rate of 2×10^{-5}, a batch size of 4, and a total of 200 training epochs; the learning rate was decayed via cosine annealing to prevent overfitting. QAM-GPDN was trained on the same hardware configuration. For each scene, 1024 grasp poses were randomly sampled, and the point cloud branch sampled 1024 points from depth data filtered by the functional region mask. This model was optimized with AdamW (initial learning rate = 1×10^{-4}, batch size =

TABLE I
EVALUATION FOR SEGMENTATION METHODS

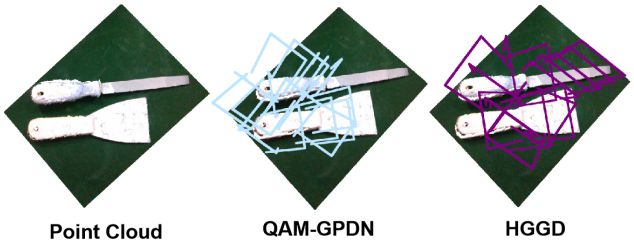
Model	**Params**	**mIoU**	
		Seen	***Unseen***
AsymFormer	33.0 M	64.6	51.7
DFormerv2-S	26.7 M	64.7	50.1
DFormerv2-B	53.9 M	66.9	55.3
MG-GRSN (Ours)	39.9 M	**70.4**	**60.5**

TABLE II
EVALUATION FOR POSE DETECTION METHODS

Model	**Seen**			**Unseen**		
	AP	$AP_{0.8}$	$AP_{0.4}$	AP	$AP_{0.8}$	$AP_{0.4}$
RGBD Grasp	30.19	37.65	24.39	28.59	36.88	20.96
GraNet	45.77	58.23	34.29	40.37	51.93	32.91
HGGD	64.93	74.58	**62.21**	53.17	64.26	45.37
QAM-GPDN (Ours)	**67.82**	**80.51**	61.39	**56.93**	**72.31**	**49.16**



(a) Functional Region Segmentation



(b) 6-DoF Grasp Detection

Fig. 5. Qualitative comparison of functional region segmentation and 6-DoF grasp detection. (a) Our MG-GRSN accurately segments tool functional regions, while baselines like DFormerv2 produce erroneous boundaries. (b) Our QAM-GPDN generates grasp proposals concentrated in functional regions, whereas methods like HGGD often propose grasps in non-functional areas.

TABLE III
ABLATION STUDIES

Method	Task	mIoU	AP
Full		**70.4**	-
SET 1	Region Segmentation	52.6	-
SET 2		64.3	-
Full		-	**67.82**
SET 3	Pose Detection	-	59.64
SET 4		-	52.38

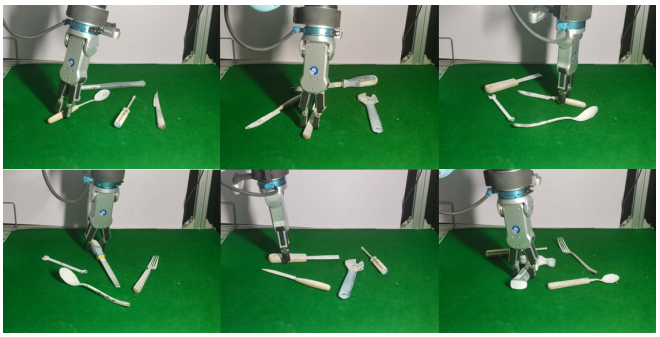
4, 200 training epochs), with cosine annealing learning rate scheduling applied.

B. Comparison with State-of-the-Art Methods

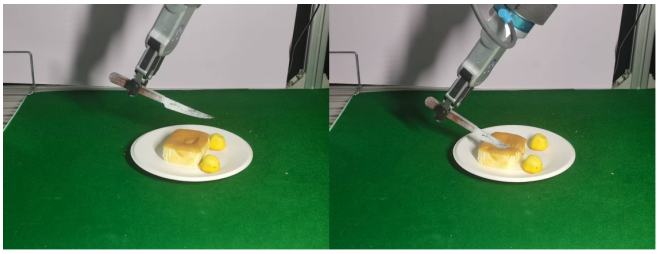
We compare Tool-Grasp against representative baselines for functional region segmentation and 6-DoF grasp detection. AsymFormer [20] and DFormerv2 [18] were adapted for part-level segmentation by fine-tuning on Tool-Grasp's functional region annotations (replacing their original instance-level labels). We retrained RGBD Grasp [22], GraNet [23], and HGGD [24] on Tool-Grasp, with pose formats converted to each method's input specifications; grasps were sampled across the entire tool surface (without prioritizing functional regions). Table I summarizes the mIoU results for functional region segmentation. Our MG-GRSN outperforms baselines by 3.5% (seen) and 5.2% (unseen) in mIoU, demonstrating its ability to model intra-object details via mask-guided feature refinement. For 6-DoF grasp detection (Table II), QAM-GPDN achieves 2.89% and 3.76% higher AP on seen and unseen tools, respectively, thanks to region-aware point cloud filtering and dynamic multi-modal fusion. Fig. 5 shows qualitative results: MG-GRSN accurately segments functional regions, and QAM-GPDN prioritizes poses within these regions. In contrast, baselines often fail to distinguish functional parts or generate valid poses outside critical regions.

TABLE IV
RESULTS OF REAL-ROBOT EXPERIMENTS

Objects ID	Type	Validity Rate	Success Rate
1,2,3,4	seen	100%	90%
5,6,7,8	seen	90%	80%
9,10,11,12	seen	90%	70%
17,18,19,20	seen	100%	80%
1,5,9,17	seen	90%	90%
13,14,15,16	unseen	90%	70%



(a) Examples of functional grasps



(b) Operation using functional grasps

Fig. 6. High-score functional grasps and their application in operation. (a) The center points of grasps for tools including knives, forks, spoons, hammers, scrapers, and files are located in functional regions, ensuring stable grasping. (b) Using high-score functional grasps enables the robot to successfully complete subsequent tool-using tasks, such as grasping a knife to cut bread.

C. Ablation Studies

To validate the contribution of key modules, we conducted ablation experiments on the seen subset of the Tool-Grasp Dataset. Results are summarized in Table III.

1) MG-GRSN Ablations: Two ablation settings were designed for MG-GRSN: (1) Removing the Dynamic Mask Guidance module (SET 1), relying solely on raw RGB-D features for segmentation; (2) Replacing the gate fusion mechanism with simple element-wise addition (SET 2).

2) QAM-GPDN Ablations: For QAM-GPDN, we tested two variants: (1) Using unfiltered point clouds (including non-functional regions) instead of mask-filtered ones (SET 3); (2) Replacing Dynamic Quality-Aware Fusion with fixed-weight concatenation for multi-modal features (SET 4).

From Table III, all ablation settings exhibited significant

performance drops compared to the full framework. This confirms that each key design in MG-GRSN and QAM-GPDN is indispensable to the framework's effectiveness.

D. Real-Robot Experiments

Experiments were conducted in cluttered scenes, where four tools were randomly selected from the Tool-Grasp Dataset. For each group, ten trials were performed with language-specified targets (e.g., "fork"). Table IV presents the validity rate (percentage of poses whose centers lie in ground-truth functional regions) and success rate (percentage of poses that are both valid and yield stable grasps). Fig. 6 shows examples of high-score functional grasps and demonstrates their potential for subsequent tool operations. These results further validated the effectiveness of our proposed Tool-Grasp framework: it not only generates valid grasp poses with centers in real functional regions but also ensures stable grasping, laying a reliable foundation for subsequent tool operations. Additionally, videos of physical experiments were uploaded to the Supplementary Material for reference.

V. CONCLUSION

This paper constructs the Tool-Grasp Dataset with part-level annotations and proposes MG-GRSN and QAM-GPDN for functional region segmentation and functional grasp detection, respectively. Experimental results showed that MG-GRSN improves mIoU by 3.5% and 5.2% for seen and unseen tools, respectively. Under the same experimental settings, QAM-GPDN increases AP by 2.89% and 3.76% for these tool categories. Real-robot experiments further verified the effectiveness of the proposed method in practical scenarios.

REFERENCES

[1] R. Detry, J. Papon and L. Matthies, "Task-oriented grasping with semantic and geometric scene understanding," *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Vancouver, BC, Canada, 2017, pp. 3266-3273.

[2] M. Aburub, K. Higashi, W. Wan and K. Harada, "Functional Eigen-Grasping Using Approach Heatmaps," in *IEEE Robotics and Automation Letters*, vol. 10, no. 4, pp. 3771-3778, April 2025.

[3] J. Chen, Y. Chen, J. Zhang and H. Wang, "Task-Oriented Dexterous Hand Pose Synthesis Using Differentiable Grasp Wrench Boundary Estimator," *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Abu Dhabi, United Arab Emirates, 2024, pp. 5281-5288.

[4] A. Depierre, E. Dellandréa and L. Chen, "Jacquard: A Large Scale Dataset for Robotic Grasp Detection," *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Madrid, Spain, 2018, pp. 3511-3516.

[5] N. Silberman, D. Hoiem, P. Kohli and R. Fergus, "Indoor Segmentation and Support Inference from RGBD Images," *2012 European Conference on Computer Vision (ECCV)*, Berlin, Heidelberg, 2012, pp.746-760.

[6] S. Song, S. P. Lichtenberg and J. Xiao, "SUN RGB-D: A RGB-D scene understanding benchmark suite," *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, 2015, pp. 567-576.

[7] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser and M. Nießner, "ScanNet: Richly-Annotated 3D Reconstructions of Indoor Scenes," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, 2017, pp. 2432-2443.

[8] H. -S. Fang, C. Wang, M. Gou and C. Lu, "GraspNet-1Billion: A Large-Scale Benchmark for General Object Grasping," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, 2020, pp. 11441-11450.

[9] J. Li and D. J. Cappelleri, "Sim-Grasp: Learning 6-DOF Grasp Policies for Cluttered Environments Using a Synthetic Benchmark," in *IEEE Robotics and Automation Letter*, vol. 9, no. 9, pp. 7645-7652, Sept. 2024.

[10] A. -L. Wang, N. Chen, K. -Y. Lin, Y. -M. Li and W. -S. Zheng, "Task-Oriented 6-DoF Grasp Pose Detection in Clutters," *2025 IEEE International Conference on Robotics and Automation (ICRA)*, Atlanta, GA, USA, 2025, pp. 5692-5698.

[11] Tjeard van Oort, Dimity Miller and Will N. Browne, "Open-Vocabulary Part-Based Grasping," *arXiv preprint arXiv:2406.05951*, 2024

[12] Z. Gao, J. Deng, Z. Wan, H. Zhang, Y. Wang and J. Hu, "GraspCLIP: Pick Up what You Want," *2025 IEEE International Conference on Industrial Technology (ICIT)*, Wuhan, China, 2025, pp. 1-6.

[13] V. Lepetit, F. Moreno-Noguer and P. Fua, "EPnP: An Accurate O(n) Solution to the PnP Problem," in *International Journal of Computer Vision*, vol. 81, pp. 155-166, Feb. 2009.

[14] A. Kirillov, "Segment Anything," *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, Paris, France, 2023, pp. 3992-4003.

[15] Z. Tan, J. Feng and J. Zhou, "SGNet: Structure-Aware Graph-Based Network for Airway Semantic Segmentation," *2021 International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, Cham, Switzerland, 2021, pp. 155-165.

[16] A. Wu and L. Fu, "DBCAN: DFormer-Based Cross-Attention Network for RGB Depth Semantic Segmentation," in *Applied Sciences*, vol. 14, no. 18, p. 8329, Sep. 2024.

[17] S. Wei, Z. Zhou, Z. Lu, Z. Yuan and B. Su, "HDBFormer: Efficient RGB-D Semantic Segmentation With a Heterogeneous Dual-Branch Framework," in *IEEE Signal Processing Letters*, vol. 32, pp. 91-95, 2025.

[18] B. -W. Yin, J. -L. Cao, M. -M. Cheng and Q. Hou, "DFormerv2: Geometry Self-Attention for RGBD Semantic Segmentation," *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Nashville, TN, USA, 2025, pp. 19345-19355.

[19] Z. Wei et al., " D(R,O) Grasp: A Unified Representation of Robot and Object Interaction for Cross-Embodiment Dexterous Grasping," *2025 IEEE International Conference on Robotics and Automation (ICRA*, Atlanta, GA, USA, 2025, pp. 4982-4988.

[20] S. Du, W. Wang, R. Guo, R. Wang and S. Tang, "AsymFormer: Asymmetrical Cross-Modal Representation Learning for Mobile Platform Real-Time RGB-D Semantic Segmentation," *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Seattle, WA, USA, 2024, pp. 7608-7615.

[21] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 2016, pp. 770-778.

[22] M. Gou, H. -S. Fang, Z. Zhu, S. Xu, C. Wang and C. Lu, "RGB Matters: Learning 7-DoF Grasp Poses on Monocular RGBD Images," *2021 IEEE International Conference on Robotics and Automation (ICRA)*, Xi'an, China, 2021, pp. 13459-13466.

[23] H. Wang, W. Niu and C. Zhuang, "GraNet: A Multi-Level Graph Network for 6-DoF Grasp Pose Generation in Cluttered Scenes," *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Detroit, MI, USA, 2023, pp. 937-943.

[24] S. Chen, W. Tang, P. Xie, W. Yang and G. Wang, "Efficient Heatmap-Guided 6-Dof Grasp Detection in Cluttered Scenes," in *IEEE Robotics and Automation Letters*, vol. 8, no. 8, pp. 4895-4902, Aug. 2023.

[25] L. -C. Chen, G. Papandreou, I. Kokkinos, K. Murphy and A. L. Yuille, "DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834-848, 1 April 2018.