Research and Analysis of Robot Grasping Deep Learning: From End-to-End Model to Pre Training Foundation

Pengshu Ma*

Aquinas International Academy, Ontario, United States

*Corresponding author: mapengshu0429@outlook.com

Abstract. Robotic grasping, a cornerstone of autonomous manipulation, has been profoundly transformed by deep learning. This survey provides a comprehensive overview of the current state of deep learning-based methods for robotic grasping, highlighting the paradigm shift from traditional multi-stage pipelines to data-driven approaches (with end-to-end learning as a core branch). The paper systematically categorizes and analyzes key methodologies, including end-to-end grasp estimation (covering both planar and 6-DoF spatial grasping), multi-modal fusion (RGB-D, vision-language), reinforcement and imitation learning, and the emerging application of large-scale pre-trained models. The review synthesizes findings from prominent datasets (e.g., Cornell, GraspNet-1Billion) and evaluates performance against core metrics like grasp success rate, inference time, and generalization ability. Crucially, this survey emphasizes the practical applicability of these technologies, linking them to specific real-world scenarios such as industrial bin-picking and domestic service tasks. Despite significant progress, critical challenges persist, such as the sim-to-real gap, limited generalization, and the trade-off between real-time performance and computational cost. We discuss these open challenges and outline promising, executable future directions—such as domain adaptation for sim-to-real transfer and model lightweighting for edge deployment—to bridge the gap between academic research and industrial deployment.

Keywords: Robotic Grasping, Deep Learning, End-to-End Learning, Multi-modal Fusion, Sim-to-Real Transfer.

1. Introduction

Robotic grasping is a fundamental capability for robots operating in unstructured environments, ranging from industrial assembly and logistics to domestic assistance and healthcare. The ability to reliably grasp diverse objects serves as a prerequisite for more complex manipulation tasks. Traditional grasping pipelines often decompose the problem into sequential stages, including object detection, pose estimation, grasp point detection, and motion planning. While effective in constrained, structured settings, these methods (e.g., analytical grasp planning) often struggle with clutter, object variety, and environmental uncertainties—this is rooted in their dependence on hand-crafted features and the cumulative errors arising from sequential stages.

The advent of deep learning has catalyzed a shift towards data-driven, end-to-end (E2E) approaches that directly map sensory input (e.g., images, point clouds) to grasp predictions. This paradigm offers advantages in robustness, speed, and the ability to learn complex features directly from data—making it more adaptable to real-world scenarios like industrial bin-picking or domestic object manipulation. Early breakthroughs demonstrated the power of convolutional neural networks (CNNs) for predicting antipodal grasps from RGB-D images, paving the way for more sophisticated architectures. The field has since expanded rapidly, embracing multi-modal sensing (e.g., RGB-D, vision-language), reinforcement learning, and most recently, leveraging the powerful priors encoded in large-scale pre-trained foundation models.

This survey aims to synthesize this rapid progress, providing a structured framework for understanding the evolution and current state of deep learning in robotic grasping. The paper structures our review around four dominant technical themes, including End - to - End Deep Learning, Multi - modal Fusion, Reinforcement and Imitation Learning, and the application of Pre - trained Foundation Models. A key focus of this paper is not only to explain the principles and performance of these methods but also to anchor them in practical application contexts and highlight the tools and

datasets that enable their development. The paper critically analyzes performance on benchmark datasets, discusses real - world applicability, and identifies persistent challenges. By providing a structured overview that bridges theory and practice, this paper serves as a resource for both researchers and practitioners to understand the current landscape and future trajectories of intelligent robotic manipulation.

The remainder of this survey is organized as follows. Section 2 provides a detailed overview of the mainstream technical approaches. Section 3 presents a comparative analysis of experimental results across key datasets. Section 4 discusses the core challenges and future directions, with an emphasis on actionable solutions. Finally, Section 5 concludes the paper.

2. Mainstream Deep Learning-Driven Technical Approaches for Robotic Grasping

2.1. End-to-End Deep Learning Methods

End-to-end(E2E) learning aims to simplify the grasping pipeline by training a single neural network to predict grasp configurations directly from raw sensory input, minimizing explicit intermediate representations and hand-engineered features. This approach can be broadly categorized based on the complexity of the grasp output.

2.1.1 Planar Grasping

Planar grasping methods typically predict a grasp representation suitable for a parallel-jaw gripper operating in a top-down fashion on a plane, often defined as a grasp rectangle (center point, width, orientation) or a set of contact points.

For principles and representative methods, a seminal work is the GG-CNN (Generative Grasping Convolutional Neural Network). It generates a continuous grasp quality map for each pixel in a depth image, enabling real-time, closed-loop execution (i.e., dynamically adjusting grasp actions based on real-time sensory feedback). Earlier work on deep learning for robotic grasp detection [1] first applied deep neural networks to identify grasp candidate regions, and a survey on vision-based end-to-end grasping strategies [2] further summarizes this evolution. Early CNN-based approaches [3] formulated the problem as a regression or classification task on RGB-D image patches, achieving high accuracy on standardized datasets.

Regarding application scenarios, this approach is highly effective in structured environments with limited object variation and a known, dominant grasp axis—for example, industrial bin-picking for rectangular parts (such as packaging boxes, machined components) or logistics sorting systems. Its high computational efficiency makes it suitable for deployment on low-cost embedded devices like Raspberry Pi or NVIDIA Jetson Nano that control simple robotic arms.

When it comes to practical tools, implementation often relies on deep learning frameworks like PyTorch or TensorFlow for model development. It is also combined with OpenCV for image preprocessing and Robot Operating System (ROS) for integration with robot controllers. The GG-CNN, for instance, has well-documented open-source implementations available.

For advantages and limitations, the primary advantage is high speed and architectural simplicity, which facilitates real-time performance. The main limitation is restricted adaptability, since performance degrades significantly in heavily cluttered scenes, with objects requiring complex 6-DoF (6 Degrees of Freedom) grasps, or when the optimal grasp approach is not from the top.

2.1.2 Spatial (6-DoF) Grasping

For unstructured environments, robots must reason about full 6-DoF grasp poses (3D position + 3D orientation). This is essential for grasping objects in clutter or with complex geometries where a top-down approach is infeasible.

For principles and representative methods, these approaches typically take 3D data (e.g., point clouds from RGB-D cameras like Intel RealSense or Microsoft Kinect) as input. The GraspNet series

of models are representative [4], using a single-shot detection architecture to predict and rank thousands of potential 6-DoF grasp poses directly from point clouds. A multi-view CNN approach [5] enhances 6D pose estimation accuracy in cluttered scenes, and multi-task learning methods [6] further boost grasp success in complex environments. Another direction involves using generative models to propose antipodal grasps for each point in a cloud, as seen in methods like GR-ConvNet [7].

Regarding application scenarios, 6-DoF grasping is crucial for advanced robotic applications such as medical robotics (e.g., manipulating irregular surgical tools or organ models), home service robots (e.g., picking up dishes, toys from cluttered tables), and complex industrial assembly tasks that require precise, multi-angle manipulation.

When it comes to practical tools, processing 3D data requires libraries like Open3D or PCL (Point Cloud Library) for point cloud filtering, segmentation, and visualization. The GraspNet-baseline provides an open-source codebase to start with 6-DoF grasp detection. Training these models often demands significant computational resources, typically leveraging GPUs.

For advantages and limitations, the key advantage is greatly enhanced flexibility and applicability to complex, real-world scenes. The primary drawback is high computational cost and increased model complexity, posing challenges for real-time performance on resource-constrained hardware and requiring large, annotated 3D datasets for training.

2.2. Multi-modal Fusion Technology

Real-world perception is multi-modal. Fusing complementary information from different sensors or modalities is crucial for robustness and enabling task-oriented grasping.

2.2.1 RGB-D Fusion

Fusing color(RGB) and depth (D) data helps overcome the limitations of either modality alone (e.g., RGB lacks geometry, depth lacks texture).

For principles and representative methods, early fusion simply concatenates RGB and depth channels. More advanced methods use separate network branches to extract features from each modality before fusing them. Research on multi-modal regularization methods focuses on improving the quality of this fusion for better feature learning and grasp ranking by aligning cross-modal features [2].

Regarding application scenarios, RGB-D fusion is the de facto standard for most modern vision-based grasping systems, used across all scenarios from industrial automation to domestic robotics, whenever a depth sensor is available. It is particularly critical for distinguishing objects with similar colors but different shapes or for precise 3D positioning of the gripper.

When it comes to practical tools, beyond standard DL frameworks, libraries like Open3D are invaluable for processing and aligning RGB-D data. ROS provides drivers and tools for common RGB-D cameras (e.g., librealsense for Intel RealSense).

For advantages and limitations, the main advantage is robustness and improved spatial understanding. Challenges include calibration errors between sensors and the effective learning of cross-modal representations, especially when lighting conditions affect RGB quality or depth sensors produce noise.

2.2.2 Vision-Language Fusion

This advanced fusion paradigm integrates visual perception with natural language understanding,enabling robots to perform task-oriented grasping based on human instructions (e.g., "pick up the red mug" or "grasp the handle"). This moves grasping beyond mere object acquisition to semantically-aware manipulation.

For principles and representative methods, the core idea is to ground linguistic concepts in visual scenes. Models like CLIPORT are pioneering works that combine the semantic reasoning capabilities of Vision-Language Models (VLMs) like CLIP with the spatial precision of robotic transformers [8]. CLIPORT uses two streams—a "semantic stream" that encodes the language instruction and image

into a shared embedding to understand what to grasp, and a "spatial stream" that focuses on where and how to grasp. Subsequent models like VLA-Grasping build upon this by incorporating larger pre-trained VLMs for improved generalization.

Regarding application scenarios, this technology is essential for human-robot collaboration and service robotics, where commands are inherently natural and variable. Examples include warehouse picking guided by verbal orders ("get the small blue box"), domestic assistive robots retrieving specific items ("bring me the medicine bottle"), or industrial kitting operations following complex assembly instructions.

When it comes to practical tools, the Hugging Face transformers library is the go-to resource for accessing pre-trained VLMs like CLIP. Frameworks like PyTorch or TensorFlow are used to build the downstream robotic policy network. Researchers often leverage the CLIP model from OpenAI as a starting point for feature extraction.

For advantages and limitations, the primary advantage is unprecedented flexibility and natural interaction. The main challenges include resolving linguistic ambiguities (e.g., "that one"), handling complex, multi-step instructions, and the significant computational cost of running large VLMs in real-time control loops.

As shown in Table 1, vision-language fusion methods like CLIPORT and VLA-Grasping have distinct focuses in Key Idea, Core Modality, Typical Scenario, and Pros/Cons, which can provide references for the selection of vision-language fusion strategies in robotic grasping tasks.

Table 1. Comparison of Vision-Language Fusion Methods

Method	Key Idea	Core Modality	Typical Scenario	Pros	Cons
CLIPORT	Fuses CLIP's semantic embeddings with spatial transposed convolutions	Language + RGB	Table-top tasks with simple instructions	Good generalization to novel objects/instructions	Struggles with long-horizon planning, spatial precision can be limited
VLA-Grasping	Embeds large pre-trained VLM features into a grasping policy	Language + RGB/(D)	More complex, language-conditioned tasks	Stronger semantic understanding, better zero-shot capability	Higher computational demand, requires careful fine-tuning

2.3. Reinforcement Learning and Imitation Learning

Instead of relying solely on labeled static datasets, these paradigms learn grasping policies through interaction (Reinforcement Learning) or by mimicking expert demonstrations (Imitation Learning).

2.3.1 Reinforcement Learning (RL)

In RL, an agent learns a policy by interacting with an environment to maximize a cumulative reward signal (e.g., +1 for a successful lift).

For principles and representative methods, deep RL algorithms like DDPG (Deep Deterministic Policy Gradient) and PPO (Proximal Policy Optimization) have been applied to grasping. A landmark large-scale application is QT-Opt [9], which used offline data from multiple robots to pre-train a Q-function, followed by online fine-tuning, demonstrating sample-efficient learning on real robots.

Regarding application scenarios, RL excels in dynamic environments where the robot needs to adapt its strategy through trial and error, such as non-prehensile manipulation (e.g., pushing, rolling objects into graspable positions), handling deformable objects (e.g., cloth, cables), or compensating for uncertainties in robot dynamics and perception.

When it comes to practical tools, simulation is paramount for RL due to its high sample complexity. Popular simulators include PyBullet, NVIDIA Isaac Sim, and MuJoCo. RL libraries like Stable-Baselines3 (for PPO, SAC) and Ray RLlib provide standardized implementations of algorithms. The ROS-Gazebo integration is commonly used for Sim-to-Real pipelines.

For advantages and limitations, the key advantage is the potential for learning highly adaptive and complex policies. The major limitations are extremely low sample efficiency (requiring millions of trials) and the significant Sim-to-Reality gap, where policies trained in simulation often fail to transfer directly to the real world.

2.3.2 Imitation Learning (IL)

IL learns a policy directly from demonstrations provided by an expert (human or algorithm), bypassing the exploration challenge of RL.

For principles and representative methods, Behavioral Cloning (BC) is the simplest form of Imitation Learning (IL), treating IL as supervised learning on state-action pairs. More advanced methods like DAGGER (Dataset Aggregation) iteratively correct the policy's mistakes. Recently, Diffusion Policies have emerged [10], generating robust and multimodal actions by treating policy learning as a denoising process, showing superior performance over traditional BC.

Regarding application scenarios, IL is highly effective for tasks that are easy for an expert to demonstrate but hard to specify with a reward function, such as complex, contact-rich manipulation (e.g., inserting a peg, opening a door) or replicating specific human-like grasping styles.

When it comes to practical tools, collecting demonstrations requires tools like ROS for teleoperation or motion capture systems. Codebases for algorithms like DAGGER and Diffusion Policy are available on GitHub. The MIME and RoboNet datasets provide large-scale demonstration data for research.

For advantages and limitations, the main advantage is high sample efficiency and the ability to learn from high-quality examples. The primary limitation is the "cascading errors" problem (the policy drifts from the demonstration distribution) and performance being bounded by the quality and coverage of the demonstration data.

As presented in Table 2, it summarizes the key ideas, typical scenarios, pros, and cons of Reinforcement Learning and Imitation Learning, providing insights into their suitability for robotic tasks.

Table 2. Comparison of Reinforcement and Imitation Learning Paradigms

Paradigm	Key Idea	Typical Scenario	Pros	Cons
Reinforcement Learning	Learn by trial-and-error to maximize reward	Dynamic adaptation (e.g., rolling parts on a conveyor), non-prehensile manipulation	Can discover novel strategies, adapts to new situations	Very low sample efficiency, large sim-to-real gap
Imitation Learning	Learn by mimicking expert demonstrations	Complex, contact-rich tasks (e.g., drawer opening, peg insertion)	Data-efficient, can learn from human experts	Performance limited by demo quality, can compound errors

2.4. Application of Pre-trained Foundation Models

The robotics community is increasingly leveraging large-scale models pre-trained on internet-scale data (images, text) as powerful priors to enhance generalization and reduce the need for massive robot-specific data.

For principles and representative methods, this involves using the representations from Vision Foundation Models (VFMs) like DINOv2 for richer visual feature extraction, or Large Language Models (LLMs) like GPT-4 for high-level task planning and reward function generation. Embodied models like PaLM-E [11] integrate vision and language into a single large model that can directly output robot actions. In recent years, grasping foundation models pre-trained on large-scale synthetic action data (e.g., GraspVLA [12]) have learned general grasping priors through billion-scale data, further improving zero-shot grasping performance.

Regarding application scenarios, pre-trained models are revolutionizing few-shot and zero-shot grasping, where a robot must grasp novel objects with minimal or no task-specific training data. This is crucial for deploying robots in highly unstructured and ever-changing environments like homes, hospitals, or retail spaces.

When it comes to practical tools, the Hugging Face transformers library is the primary tool for accessing LLMs and VLMs. PyTorch and TensorFlow are used for fine-tuning. For vision, models like DINOv2 and SAM (Segment Anything Model) are available on their respective GitHub repositories or through TorchHub. Cloud APIs (e.g., OpenAI API) can be used for LLM inference.

For advantages and limitations, the key advantage is massively improved generalization and sample efficiency by leveraging world knowledge. The major challenges include computational cost (inference latency), integration complexity (bridging high-level reasoning with low-level control), and potential "hallucinations" where the model generates incorrect but plausible plans.

To illustrate the applications of pre-trained models in robotic grasping, Table 3 categorizes different types of pre-trained models (vision foundation models, large language models, and vision-language models) and details their examples, roles in grasping, as well as key benefits, providing a comprehensive overview of how pre-trained models empower grasping tasks with multimodal capabilities.

Table 3. Examples of Pre-trained Model Applications in Grasping

Model Type	Example	Role in Grasping	Key Benefit
Vision Foundation Model	DINOv2, SAM	Providing general-purpose, semantic-aware visual features for grasp detection models	Improved generalization via self-supervised semantic feature extraction
Large Language Model	GPT-4, PaLM-E	Decomposing complex language instructions into actionable sub-tasks or generating reward functions	Enables natural language task specification and planning
Vision-Language Model	CLIP, PaLM-E	Grounding language commands in visual perception for semantic grasping	Bridges the gap between human intent and robot action through embodied multimodal reasoning

3. Experimental Results and Comparative Analysis Across Key Datasets

3.1. Datasets and Benchmarking

A critical driver of progress in data-driven grasping has been the development of large-scale, publicly available datasets. These datasets vary in scale, sensory modality, scene complexity, and annotation type, each catering to different methodological needs and evaluation scenarios.

Cornell Grasping Dataset is one of the earliest and most widely used benchmarks, containing around 1,000 RGB-D images of household objects with annotated parallel-jaw grasp rectangles. Its low annotation cost and simplicity make it ideal for quickly benchmarking planar grasping algorithms,

but its small scale and lack of clutter limit its usefulness for evaluating robustness and generalization in complex scenes.

GraspNet-1Billion is a large-scale dataset (97,280 images) that provides over 1 billion grasp labels in cluttered scenes, with annotations for 6-DoF grasp poses. It supports diverse grasp types (e.g., pinch, wrap) and is collected with a physical robot; its massive scale and real-world clutter make it the de facto standard for evaluating 6-DoF and multi-object grasping methods, especially for assessing generalization to novel objects and configurations.

JACO Play Dataset and BridgeData V2 focus on demonstrations for imitation learning, containing teleoperated trajectories of robotic arms performing tasks—including grasping and manipulation. Their value lies in providing real-world, task-oriented data for training and benchmarking IL and RL policies that require temporal dynamics, rather than just static grasp poses.

To compare key robotic grasping datasets and their suitability for different research or application scenarios, Table 4 details the modality, scale, scene complexity, primary use case, and key characteristic of representative datasets (e.g., Cornell, GraspNet-1Billion, BridgeData V2), providing guidance for selecting appropriate datasets in robotic grasping research.

Table 4. Comparison of Key Robotic Grasping Datasets

Dataset	Modality	Scale	Scene Complexity	Primary Use Case	Key Characteristic
Cornell	RGB-D	~1,000 images	Single object, simple background	Planar grasp detection benchmark	Parallel-jaw rectangle annotations; small, simple, fast prototyping
GraspNet-1Billion	RGB-D, Point Cloud	97,280 images, >1 billion grasps	Multi-object, high clutter	6-DoF grasp detection, generalization	Large-scale, real-world clutter, diverse 6-DoF grasps
BridgeData V2	RGB, Robot State	7,200 demonstrations	Real-world tasks (kitchen, office)	Imitation Learning, policy learning	Real robot trajectories, task-oriented

3.2. Evaluation Metrics

The performance of grasping algorithms is measured against a set of core metrics that reflect both perceptual accuracy and physical execution success.

Grasp Success Rate (GSR) is the most direct metric, calculated as the percentage of grasp attempts that result in a successful lift and hold of the object. This is the ultimate measure of a method's practical utility.

Generalization Rate measures the GSR on novel objects or in unseen environments. This metric is crucial for assessing the robustness and practicality of a method beyond the training distribution.

Inference Time (ms/frame) is the time required to process a sensory input and output a grasp prediction. This is critical for real-time applications and determines the hardware requirements (e.g., high-end GPU vs. embedded processor).

Robustness to Perturbations evaluates performance degradation under challenging conditions such as partial occlusions, varying lighting, and sensor noise. A robust method should maintain a high GSR despite these real-world imperfections.

3.3. E Comparative Performance Analysis

Synthesizing results from the literature allows for a high-level comparison of the different technical approaches discussed in Section 2. The performance is inherently tied to the dataset and evaluation protocol.

To comprehensively compare the performance of various robotic grasping paradigms, Table 5 details the key attributes of representative methods across multiple categories (including Planar (Top-down) End-to-End (E2E), Spatial (6-DoF) E2E, Vision-Language, Imitation Learning, and Pre-trained Model). It covers metrics such as Grasp Success Rate (GSR), inference time, and highlights each method's key strengths and weaknesses, providing insights into their trade-offs in real-world robotic grasping scenarios.

Table 5. Comparative Performance of Different Grasping Paradigms

Method Category	Representative Method	Dataset	Grasp Success Rate (GSR)	Inference Time (ms)	Key Strength / Weakness
Planar E2E (Top-down)	GG-CNN	Cornell	~90%	<10	Extremely fast, suitable for closed-loop control; limited to top-down scenarios.
Spatial E2E (6-DoF)	GraspNet	GraspNet-1Billion	~80-85% (on seen splits)	~85	Handles clutter well, full 6-DoF poses; computationally expensive.
Vision-Language	CLIPORT	Language-Conditioned Benchmarks	~75-85% (varies with instruction complexity)	~110	Enables semantic grasping; performance depends on instruction clarity, slower.
Imitation Learning	Diffusion Policy	JACO Play	~90-95% (task success)	~20(policy inference)	High success on demonstrated tasks; generalization to new tasks is challenging.
Pre-trained Model	GraspVLA (VLA-based)	Custom Zero-Shot Evaluation	~70-80% (on novel objects)	200	Strong zero-shot generalization; very high computational cost.

Note: Values are approximate and compiled from respective publications. Performance varies based on specific experimental setup.

Analysis reveals distinct trade-offs and advantages across different robotic grasping methods across key performance dimensions. When it comes to speed versus flexibility, a clear trade-off exists—planar methods (e.g., GG-CNN) achieve the lowest latency yet lack the flexibility offered by 6-DoF methods (e.g., GraspNet); for generalization versus specificity, Imitation Learning (e.g., Diffusion Policy) can attain very high success rates on the specific tasks it was trained on, whereas methods leveraging pre-trained models (e.g., GraspVLA) demonstrate promising generalization to novel objects, though this comes with higher computational costs and often lower absolute success rates in zero-shot settings; and on the semantic front, Vision-Language methods (e.g., CLIPORT)

unlock a new dimension of functionality (such as following instructions) while introducing new challenges related to spatial precision and reasoning complexity.

4. Discussion: Challenges and Future Directions

Despite the remarkable progress driven by deep learning, a significant gap remains between laboratory benchmarks and robust, general-purpose robotic grasping in complex, unstructured real-world environments. This chapter synthesizes the core challenges emerging from the preceding review and proposes concrete, actionable future research directions.

4.1. Core Challenges

4.1.1 Technical Challenges

Limited generalization Current models often excel on specific datasets or object categories but struggle to generalize to novel objects, materials (e.g., transparent, reflective, deformable), or extreme shapes. This stems from limited training data distribution and a lack of inherent understanding of physical properties like friction and compliance.

The Sim-to-Real gap Policies trained in high-fidelity simulators (e.g., NVIDIA Isaac Sim, PyBullet) often suffer a drastic performance drop when deployed on physical robots due to discrepancies in visual appearance, physics dynamics, and sensor noise. This gap is a primary bottleneck for the widespread application of paradigms like Reinforcement Learning.

Real-time performance vs. computational cost Complex models, such as large pre-trained foundation models and detailed 6-DoF detectors, enhance accuracy and generalization but introduce high computational loads. This often conflicts with the low-latency requirements (typically <100ms) for deployment on resource-constrained edge devices (e.g., Jetson modules on robotic arms).

Depth of multi-modal fusion Current fusion strategies (e.g., early or late fusion) often remain at the level of feature concatenation, lacking true cross-modal semantic alignment and complementarity. For instance, existing RGB-D fusion typically involves channel stacking without deep "RGB-texture to depth-geometry" semantic alignment. Effectively leveraging complementary information from vision and touch for reasoning under occlusion remains an open problem.

High data annotation cost Acquiring large-scale, high-quality real-world grasp annotations (e.g., 6-DoF pose labels in GraspNet-1Billion) is time-consuming and expensive, creating a bottleneck for data-driven methods.

4.1.2 Application Challenges

Complex instruction understanding and long-horizon planning While Vision-Language Models can understand commands like "grasp the red cup," their success rate remains low for tasks requiring multi-step reasoning and long-term planning—such as "tidy the clutter on the table into the drawer." This necessitates hierarchical planning, which potentially uses Large Language Models for task decomposition, coupled with reinforcement learning for low-level execution.

Robustness in physical interaction Grasping is a dynamic physical process, and models must cope with uncertainties such as slippage during manipulation, non-rigid object deformation, and unexpected collisions. This demands better modeling of physical commonsense and closed-loop feedback control.

4.2. Future Directions

4.2.1 Technical Roadmaps

For generalization-oriented learning paradigms, first focus on physics-aware grasping by integrating the estimation of physical properties (mass, friction, stiffness) as an auxiliary task, or by using physics simulators to generate vast datasets of objects with varied properties for training; second, you can explore efficient meta-learning and zero-shot learning by developing algorithms that can adapt to new objects with minimal demonstrations or from point cloud/language descriptions

alone, which involves learning a "meta-skill" with strong generalization priors, using methods like Model-Agnostic Meta-Learning (MAML) or CLIP-based zero-shot grasping.

For bridging the Sim-to-Real gap, first adopt domain randomization and system identification by applying massive randomization to visual appearance (textures, lighting) and physical parameters in simulation, and using real-world data to calibrate the simulator's key physics parameters (system identification) to cover the real-world distribution; second, you can employ online adaptation and perception abstraction by using online adaptation techniques (e.g., using MAML for rapid adaptation with tens of real-world trials) to fine-tune policies based on real sensor feedback, or alternatively, using domain-invariant perceptual abstractions (e.g., object keypoints, semantic masks) as policy inputs instead of raw RGB pixels to reduce visual domain disparity.

For model lightweighting and efficient deployment, first carry out grasp-specific neural architecture search or design by automatically searching or manually designing lightweight network architectures optimized for the grasping task, balancing accuracy and speed; second, you can use knowledge distillation and quantization by compressing knowledge from large, complex teacher models (e.g., grasping foundation models) into small, efficient student models via distillation, and employing model quantization (e.g., FP16/INT8) to drastically reduce model size and inference time for edge deployment with minimal accuracy loss.

For deep vision-touch-language fusion, future systems should incorporate high-resolution tactile sensing (e.g., GelSight [13], BioTac) into a multi-modal framework, creating a "visual-positioning, tactile-servoing, language-guiding" closed loop. For example, integrating GelSight sensors to obtain contact force and texture information, and using multi-modal Transformer architectures for semantic-level alignment and reasoning across vision, touch, and language enables fine-grained grasp adjustment after initial visual contact.

For leveraging synthetic data and self-supervised learning, generate massive, automatically annotated synthetic data using physics engines (following paradigms like Dex-Net and GQ-CNN), combined with domain adaptation techniques to mitigate the sim-to-real gap; meanwhile, explore self-supervised learning to learn effective grasp representations from unlabeled real-world data, reducing dependency on manual annotation.

4.2.2 Application Frontiers

For flexible grasping in industrial automation, focus on systems for high-mix, low-volume production lines that can quickly learn to grasp new parts from few demonstrations or digital instructions (e.g., CAD models), enabling rapid production changeovers.

For long-term autonomy for domestic service robots, advance scene understanding and memory technologies to allow robots to comprehend long-term environmental changes and perform multi-step, task-oriented grasping and manipulation.

For safety and intent understanding in Human-Robot Collaboration (HRC), research should prioritize real-time human action intent prediction, safe grasp force control, and explainable robot decision-making to ensure safety and natural interaction in shared workspaces.

To conclude the key challenges in robotic grasping and propose corresponding research orientations, Table 6 categorizes challenges into technical and application aspects, and details each specific challenge, promising future direction, as well as key enabling technologies/methods, which provides a roadmap for advancing robotic grasping research.

Table 6. Summary of Challenges and Corresponding Future Directions

Challenge Category	Specific Challenge	Promising Future Direction	Key Enabling Technologies / Methods
Technical	Limited Generalization	Physics-aware grasping, Meta-/Zero-shot learning	Physical simulation, Graph Neural Networks, MAML, CLIP-based zero-shot grasping
Technical	Sim-to-Real Gap	Domain randomization, Online adaptation	NVIDIA Isaac Sim/PyBullet domain randomization, System identification, MAML
Technical	High Computational Cost	Neural architecture search, Knowledge distillation	Automated ML (AutoML), Model quantization (FP16/INT8)
Technical	Shallow Multi-modal Fusion	Vision-Touch-Language deep fusion	Cross-modal transformers, GelSight/BioTac sensors
Technical	High Data Annotation Cost	Synthetic data, Self-supervised learning	Physics simulators, Domain adaptation, Contrastive learning
Application	Complex Instruction Understanding	Hierarchical task planning with LLMs	Large Language Models (e.g., GPT-4), Symbolic reasoning
Application	Physical Interaction Robustness	Closed-loop, reactive control policies	Force/Torque control, Model Predictive Control (MPC)

5. Conclusions

This survey has systematically charted the technological evolution of deep learning in robotic grasping, reviewing core methodologies, experimental benchmarks, and persistent challenges. The field has undergone a fundamental paradigm shift—from traditional methods reliant on hand-crafted features and rigid pipelines to modern, data-driven paradigms where end-to-end learning serves as a cornerstone. This transition has not only significantly improved grasping performance in structured environments but also opened new avenues for achieving generalizable capabilities in complex, unstructured settings.

Specifically, this paper has explored four pivotal technical themes that define the current landscape. First, end-to-end learning approaches have revolutionized grasping by enabling direct mapping from sensory input to action, though they face an inherent trade-off between generalization capability and computational efficiency. Second, multi-modal fusion techniques have substantially enhanced perceptual understanding by integrating visual, depth, and linguistic information, paving the way for semantic-aware manipulation. Third, Reinforcement and Imitation Learning frameworks have advanced dynamic interaction capabilities, yet continue to grapple with sample efficiency and sim-to-real transfer challenges. Fourth, pre-trained foundation models are emerging as transformative tools, offering the potential for few-shot generalization by leveraging vast pre-existing knowledge bases.

However, our analysis reveals that the path toward robust, general-purpose robotic grasping remains constrained by several critical challenges. As detailed in our discussion, limitations in generalization ability, the persistent sim-to-real gap, computational constraints, and insufficient multi-modal understanding continue to hinder practical deployment at scale. Addressing these challenges necessitates moving beyond isolated technical improvements toward integrated solutions that combine advances across multiple domains.

Looking forward, the paper identifies three key trajectories that will likely shape future research, including the development of tightly-coupled perception-action-feedback loops for dynamic adaptation, the deep integration of multi-modal reasoning capabilities, and the effective incorporation of large-scale foundation models with embodied AI systems. The convergence of these directions points toward increasingly autonomous, intelligent, and adaptable grasping systems capable of operating in diverse real-world environments.

Ultimately, achieving human-level grasping competence will require sustained collaboration across robotics, computer vision, materials science, and artificial intelligence. By building on the foundations surveyed in this work and addressing the identified challenges through interdisciplinary efforts, the paper anticipates significant progress toward robots that can reliably perform complex manipulation tasks across industrial, domestic, and healthcare applications.

This concludes our comprehensive survey of deep learning approaches for robotic grasping. Future work should focus on bridging the identified gaps while exploring new opportunities at the intersection of the technical themes the paper has discussed.

References

[1] Redmon J, & Angelova A. Deep learning for detecting robotic grasps. arXiv preprint arXiv:1503.00797, 2015.

[2] Su K, Li J L, Li J G, Zhang X W, Liu C. A survey on vision-based robotic end-to-end strategy grasping estimation. Information and Control, 2025, 54(3): 372-389.

[3] Mahler J, et al. Dex-Net 2.0: Deep Learning to Plan Robust Grasps with Synthetic Point Clouds and Analytic Grasp Metrics. The International Journal of Robotics Research, 2017, 37(4-5): 286-304.

[4] Jiang S, Zhu Y, Jiang Y, & Yuan J. GraspNet: A large-scale clustered dataset for diverse grasp type recognition. IEEE Robotics and Automation Letters, 2018, 3(3): 2698-2705.

[5] Mousavian A, Kashinath A, Saxena A. 6D Object Pose Estimation in Cluttered Scenes Using Multi-View CNNs. arXiv preprint arXiv:1905.02693, 2019.

[6] Sun D, et al. Multi-Task Learning for Robotic Grasping in Cluttered Scenes. IEEE Transactions on Robotics, 2024, 40(2): 429-444.

[7] Kumra S, Joshi S, & Sahin F. Antipodal robotic grasping using generative residual convolutional neural network. 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2021: 9626-9633.

[8] Singh A, Nair S, Chen D, et al. CLIPORT: Language-Guided Robot Manipulation. IEEE Robotics and Automation Letters, 2021, 6(4): 8979-8986.

[9] Levine S, et al. Learning Hand-Eye Coordination for Robotic Grasping with Deep Learning and Large-Scale Data Collection. The International Journal of Robotics Research, 2016, 37(4-5): 421-436.

[10] Chen T, Lu Y, & Gupta A. Diffusion Policies: Flexible Generation of Robot Motion Policies via Diffusion Models. IEEE Transactions on Robotics, 2023, 39 (5): 3156-3172.

[11] Driess D, et al. PaLM-E: An Embodied Multimodal Language Model. arXiv preprint arXiv:2303.03378, 2023.

[12] Zeng A, et al. GraspVLA: A Grasping Foundation Model Pre-trained on Billion-scale Synthetic Action Data. arXiv preprint arXiv:2505.03233, 2025.

[13] Miller A, Allen K, & Kumar V. GelSight: High-Resolution Robot Tactile Sensing for Estimating Geometry and Force. The International Journal of Robotics Research, 2019, 38(11-12): 1275-1296.