Generative Deep Learning Models for Grasp Pose Detection under Occluded and Partial Point Clouds: A Comprehensive Review

Anushka Singhania
Automation and Robotics
Defence Institute of Advance Technology(DU)
Pune, India
singhaniaanu2002@gmail.com

Atul Chavan
Research And Development Establishment Engineers
Pune, India
atul.chavan86@gov.in

RK Satapathy
Automation and Robotics
Defence Institute of Advance Technology(DU)
Pune, India
rksatapathy@diat.ac.in

doi: https://doi.org/10.37285/bsp.SACAD2025.67

***Abstract*-Grasp pose detection plays a central role in enabling robots to interact intelligently with unstructured environments. Recent advances in deep learning have significantly improved the accuracy and adaptability of grasp prediction models, especially when using 3D point clouds as input. However, challenges such as occlusion, domain shift, and limited real-world data continue to constrain performance. This review paper examines recent developments in generative deep learning frameworks for grasp pose detection, emphasizing methods that leverage diffusion models, transformers, and latent representation learning to generate robust grasp configurations. It systematically compares existing frameworks, datasets, and evaluation strategies, while mapping core challenges to their emerging solutions such as domain adaptation, self-supervised learning, and efficient model design. It outlines future research directions focused on bridging the gap between simulation and reality, paving the way for robotic grasping systems that are both resilient and adaptive in real world scenarios.**

Keywords- Generative Grasp Learning, Deep Learning, Point Clouds, Occlusion Robustness, Robotic Manipulation, Sim-to-Real Transfer

I. INTRODUCTION

Robotic grasping is a fundamental capability for autonomous systems operating in human-centric environments, yet it remains one of the most persistent challenges in robotics [1], [2]. Despite notable progress in structured laboratory conditions, robots still lag behind human dexterity and adaptability when dealing with cluttered and dynamic real-world settings [1]. A primary bottleneck arises when perception systems encounter occluded or partial point clouds, which are common in practical scenarios. Traditional analytical methods based on force-closure and geometric reasoning require complete object models and fully observable scenes, assumptions rarely satisfied in unstructured environments[3]. These methods, though mathematically grounded, fail to handle uncertainty and incomplete sensory data [1].

Recent advances in deep learning have significantly reshaped robotic grasp detection by enabling data-driven mapping from perception to action without explicit modelling [4]. However, conventional discriminative networks struggle to generalize under partial observations, suffer from mode collapse, and often produce a single grasp hypothesis per object, neglecting the inherently multi-modal nature of feasible grasps [5], [6]. In contrast, generative models such as Variational Autoencoders (VAEs), Generative Adversarial Networks (GANs), and Diffusion Models offer a principled framework for capturing grasp uncertainty and diversity. They can infer missing geometry, generate multiple feasible grasp poses, and adapt to unseen object shapes using partial or noisy point cloud data [7], [8].

The demand for such robust grasp generation is particularly evident in industrial and service robotics, where systems must handle diverse, unstructured objects without exhaustive reprogramming [2], [9]. Generative models thus bridge the gap between perception uncertainty and grasp reliability by sampling from learned multi-modal distributions, enabling robots to operate more autonomously across domains.

This review focuses on generative deep learning frameworks for grasp pose detection under occluded and partial point cloud conditions. It examines three major aspects: occlusion handling through completion and fusion strategies [10], [11]; grasp synthesis using VAEs, GANs, diffusion, and flow-based models [6], [12]; and comparative evaluation across datasets, architectures, and performance metrics. The paper further highlights emerging research directions, including large-scale pre-trained foundation models, vision–language conditioning, few-shot adaptation, and standardization of evaluation protocols [13].

This survey contributes a unified taxonomy of generative grasping methods, comparative dataset and performance analysis, and a synthesis of current trends and open challenges.

By consolidating these insights, it aims to serve as a comprehensive reference and roadmap for future research in generative grasp pose detection under partial and occluded perception.

II. BACKGROUND AND FUNDAMENTALS

Robotic grasping fundamentally involves perceiving an object, determining feasible grasp configurations, and executing a stable manipulation action [1], [2]. The task requires understanding both object geometry and physical interaction constraints such as friction, force closure, and grasp stability [3]. Traditional approaches rely on analytical models that compute contact points and wrenches based on complete 3D object meshes. While effective under ideal perception, these methods degrade under real-world conditions characterized by sensor noise, occlusions, and incomplete point clouds [4]. With the rise of learning-based methods, perception and grasp planning have become increasingly data-driven. Early deep learning approaches formulated grasp detection as a discriminative task, mapping RGB-D or point cloud inputs to grasp parameters or grasp quality maps using convolutional or point-based neural networks [5]. However, these methods are limited to deterministic predictions and struggle to generalize across novel objects or unseen viewpoints particularly under occluded or partial observations [7].

A grasp pose can be represented in 6 degrees of freedom (6-DOF), typically as a transformation in SE(3) space combining a 3D translation and rotation. Robust grasp estimation from point clouds unordered sets of 3D points requires networks that can encode permutation-invariant features. Architectures such as PointNet and PointNet++ introduced this capability by learning global and local geometric descriptors directly from raw point clouds [7]. These foundations later enabled generative extensions that infer grasp candidates even from incomplete object views [9].

Generative deep learning offers a probabilistic framework to model grasp diversity and uncertainty. Unlike discriminative systems that output a single deterministic grasp, generative models including Variational Autoencoders (VAEs), Generative Adversarial Networks (GANs), and Diffusion Models learn the underlying grasp distribution conditioned on partial sensory inputs [6], [10]. This allows them to synthesize multiple feasible grasp hypotheses, complete missing geometry, and reason about object affordances under occlusion. Recent research has further extended generative grasping through multimodal integration, combining vision, tactile, and language cues for context-aware manipulation [11]. Moreover, simulation-to-real transfer and domain randomization have become standard practices to mitigate dataset bias and ensure robustness in real-world deployments [12].

III. GENERATIVE MODEL FRAMEWORKS FOR GRASP POSE DETECTION

Generative deep learning models enable robotic systems to predict feasible 6-DOF grasp poses directly from incomplete or occluded point clouds. Unlike discriminative methods that map input features to a single deterministic grasp output, generative frameworks learn the underlying grasp distribution, allowing sampling of multiple physically plausible grasps for a given scene [7]. These methods can be broadly categorized into Variational Autoencoder (VAE)-based, Generative Adversarial Network (GAN)-based, Diffusion-based, and Flow-based frameworks, each balancing diversity, physical feasibility, and computational efficiency differently. Fig.1. provides a visual summary of the major generative model frameworks and how they relate to grasp-pose generation from partial/occluded point-clouds.

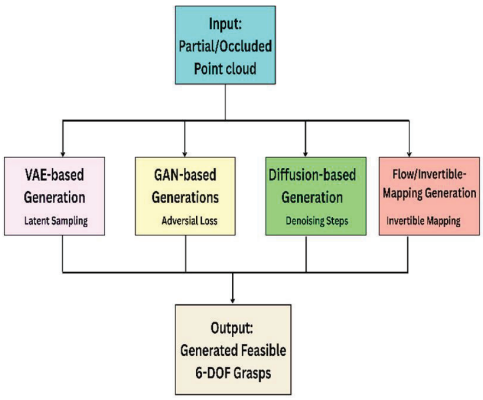

Fig.1. Generative Model Taxonomy Overview

A. VAE-Based Frameworks

VAE-based grasp generators learn a probabilistic latent space representing feasible grasp configurations. They encode partial or completed point clouds into compact latent embeddings, from which grasp poses are sampled and decoded. These models excel at modeling uncertainty under occlusion and at generating diverse grasp candidates, though they can produce blurred or averaged outputs due to latent regularization. Representative examples include ZeroGrasp and VariGrasp, which integrate kinematic feasibility constraints into their loss formulation to ensure physically valid samples [14]. As shown in Fig.2.

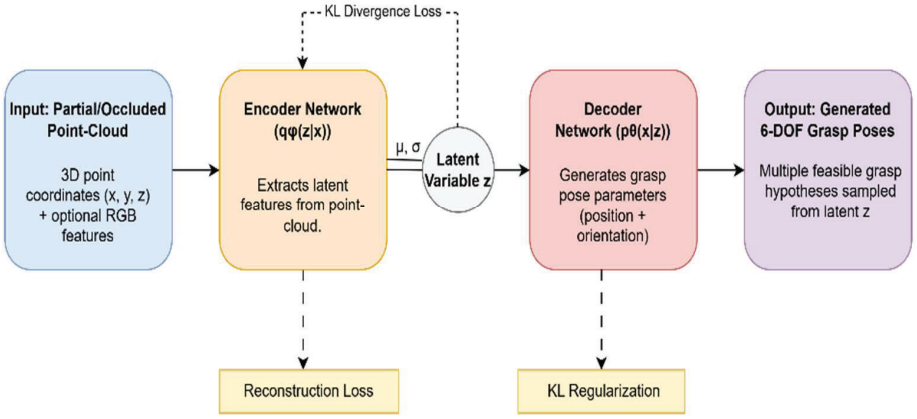

Fig.2. VAE-Based Grasp Generation Flow

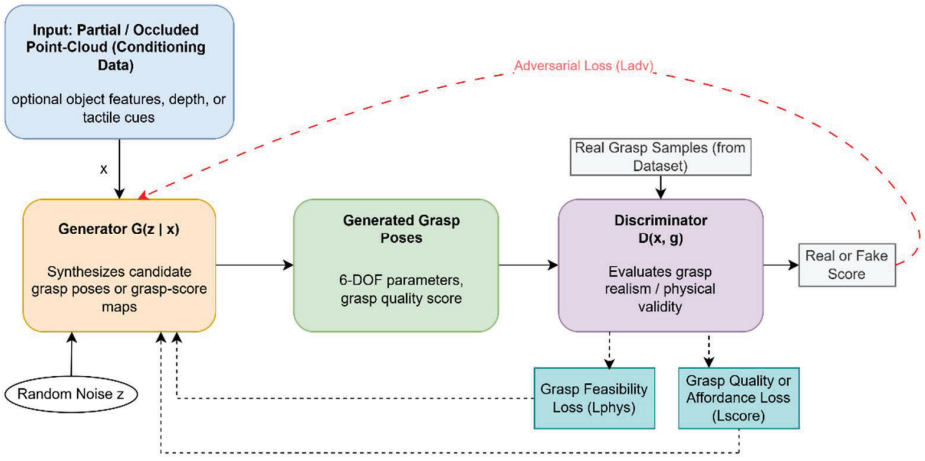

Fig.3. GAN-Based Grasp Generation Framework

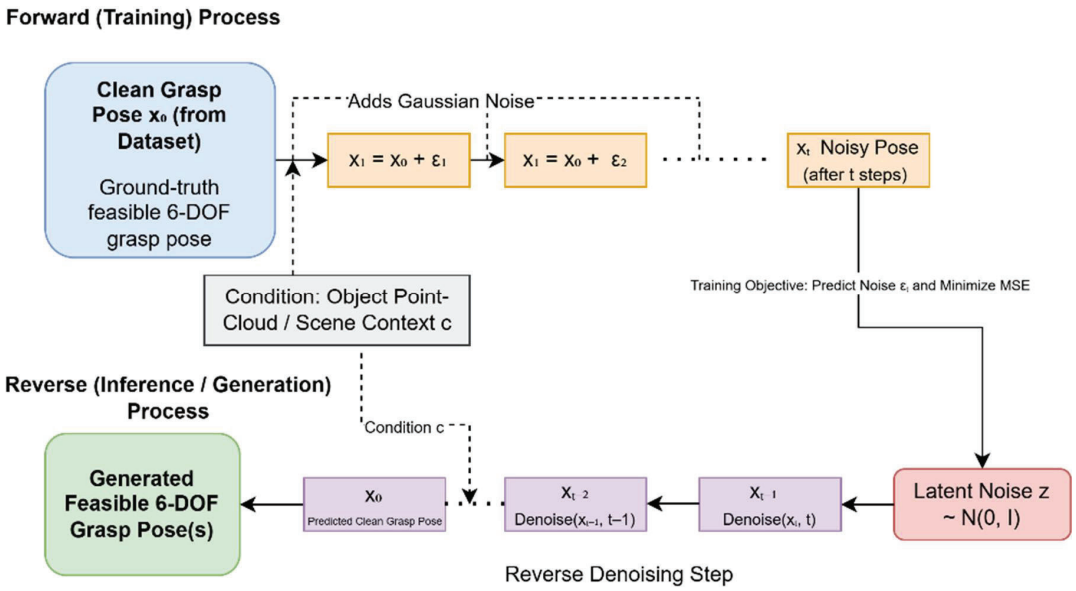

Fig.4. Diffusion-Based Grasp Generation Framework

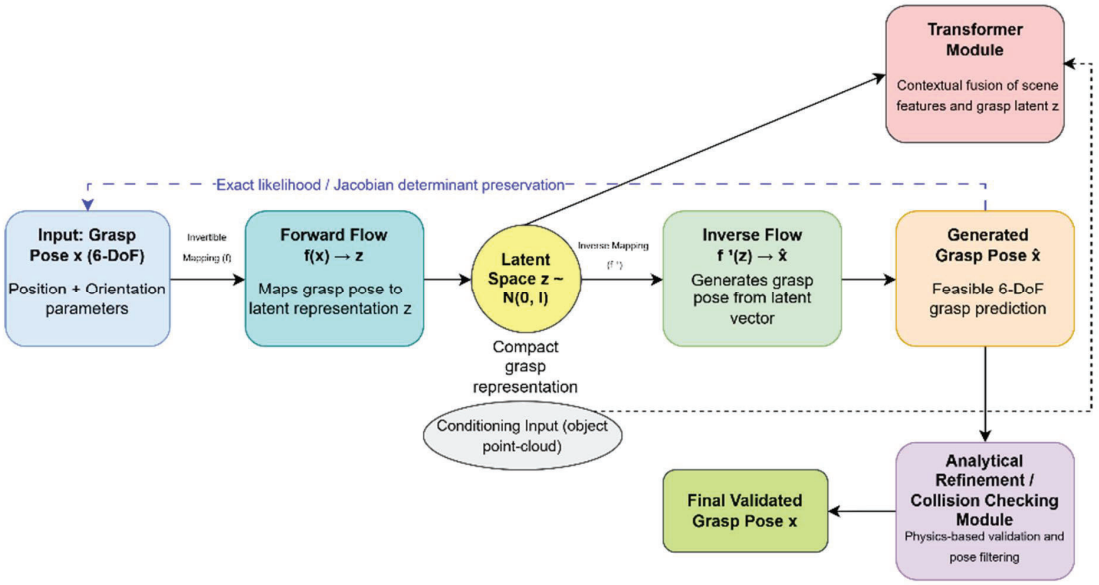

Fig.5. Flow-Based / Hybrid Generative Grasping Framework

TABLE 1: EPRESENTATIVE GENERATIVE GRASPING APPROACHES

Method	Architecture	Dataset	Input Representation	Output Representation	Performance (Success Rate/Metric)
Grasp Diffusion Network	Diffusion Model (DDPM in SO(3) × R^3)	ACRONYM (CAT10 subset, 567 objects, 1.1M grasps)	Partial Point Cloud	6-DOF Grasp (parallel-jaw)	90% (sim), 90% (real, best config)
DexDiffuser	Diffusion Model (DexSampler + DexEvaluator)	DexGraspNet (1.7M grasps, 5378 objects)	Partial Point Cloud	16-DOF Grasp (Allegro Hand)	98.77% (sim), 68.89% (real)
Dexterous Grasp Transformer (DGTR)	Transformer-based (DETR-like)	DexGraspNet	Complete Point Cloud	22-DOF Grasp (Shadow Hand)	98.77% (sim), 66.6% (real)
GraspGen	Diffusion-Transformer (DiT)	GraspGen Dataset (53M grasps, 36K objects)	Partial/Complete Point Cloud	6-DOF Grasp (multi-gripper)	65.3% (FetchBench task success), 75% (FetchBench grasp success)
Contact-GraspNet	Contact-based CNN	GraspNet-1Billion	Scene Point Cloud	6-DOF Grasp	26.1% (FetchBench task success)
Efficient Heatmap-Guided Grasp	Heatmap-based CNN	GraspNet-1Billion, TS-ACRONYM	RGB-D Image	6-DOF Grasp	28 ms inference, 64.45% AP (GraspNet-1B seen)
Constrained 6-DoF Grasp (CGDF)	Diffusion Model + Convolutional Plane Features	DA2 Dataset	Point Cloud (full/partial)	6-DOF Grasp (dual-arm)	60.3% GSR (sim), 43.51% FC (unconstrained)
GraspNet-1Billion Baseline	PointNet++ + Grasp Sampling	GraspNet-1Billion	Scene Point Cloud	6-DOF Grasp	37.7% AP (seen), 19.7% AP (similar), 9.6% AP (novel)

B. GAN-Based Frameworks

As shown in Fig.3. GAN-based methods employ an adversarial setup with a generator producing candidate grasps and a discriminator enforcing grasp realism. The generator learns to synthesize plausible poses or quality maps, while the discriminator penalizes physically implausible outputs. GANs have demonstrated strong grasp diversity and sharper pose representations, though training instability and mode collapse remain concerns [15]. Notable examples such as GraspGAN and AffordanceGAN incorporate physical plausibility or grasp-score losses, yielding high success rates on cluttered-scene datasets like GraspNet-1Billion [11].

C. Diffusion-Based Frameworks

Diffusion models, such as DexDiffuser, Grasp Diffusion Network, and GraspGen, represent the current state-of-the-art in generative grasp synthesis. They iteratively refine random noise into physically valid grasp configurations through denoising steps [13]. These models achieve superior robustness under occlusion and generalize effectively to unseen objects. However, their iterative sampling makes them computationally expensive. Diffusion-Transformer hybrids further enhance spatial reasoning by combining diffusion sampling with attention-based latent refinement [16]. As illustrated in Fig.4.

D. Flow-Based and Hybrid Frameworks

As depicted in Fig.5. Flow-based models learn invertible mappings between data and latent spaces, enabling exact likelihood estimation and efficient sampling. Though less common, they offer strong interpretability and real-time potential. Hybrid frameworks combine generative modules with analytical or regression based components for example, using diffusion for hypothesis generation followed by a physics-based grasp evaluator. Such systems balance diversity with stability and are effective in sim-to-real deployment [6].

E. Taxonomy Summary

Table 1 summarizes the main generative framework categories, their input/output representations, advantages, and limitations.

IV. HANDLING OCCLUSION AND PARTIAL DATA

Grasping under occluded or incomplete observations poses significant challenges for robotic perception. Generative frameworks mitigate these challenges through three complementary strategies point cloud completion, occlusion-aware inference, and multi-view or active perception. Together, these approaches enable reliable grasp prediction from partial point clouds, as commonly encountered in cluttered real-world scenes. Fig.6. shows Occlusion Handling Strategies in Grasp Pose Detection.

A. Point Cloud Completion Methods

Point cloud completion networks reconstruct missing geometry from partial observations, providing a denser and more complete surface for downstream grasp synthesis. The Point Completion Network (PCN) employs an encoder decoder structure with skip connections and coarse-to-fine reconstruction to generate high-fidelity object completions [12]. GRNet improves upon this approach through a grid-based folding mechanism that models detailed local geometry, achieving better performance in heavily occluded regions [17]. Probabilistic methods such as PFNet predict distributions of possible object shapes conditioned on the observed input, allowing generative grasping frameworks to account for uncertainty during grasp generation [2]. Additionally, models like Generalizing 6-DoF Grasp Detection leverage learned shape priors from synthetic datasets such as ShapeNet to jointly perform shape completion and grasp prediction, enhancing generalization to novel objects [18]. These completion-based techniques improve grasp robustness by providing fuller geometric information, reducing ambiguity in contact region estimation.

B. Occlusion-Aware Architectures

Occlusion-aware grasping networks explicitly model visibility constraints and uncertainty to improve grasp prediction from partial views. Attention-based approaches such as EAGA-Net use spatial temporal attention to emphasize visible and geometrically relevant regions during grasp inference [19]. Similarly, serialization-attention frameworks (e.g. *High-Performance Grasp Pose Detection via Point Cloud Serialization-Attention*) fuse global context with fine-grained local features, enhancing occlusion tolerance [20]. More recent methods integrate latent diffusion into the grasp generation pipeline, where partial point clouds are encoded into a compact latent space, refined through diffusion steps, and decoded into completed geometry and grasp candidates simultaneously [7]. Transformer-based models, including Constrained 6-DoF Grasp Generation, incorporate pre-trained encoders that provide shape priors, improving performance on novel or heavily occluded objects [21]. These architectures represent a shift toward uncertainty-aware generative grasping systems capable of consistent predictions even under incomplete sensory input.

C. Multi-view and Active Perception

Multi-view and active perception methods enhance grasp success by increasing the effective visibility of target objects. In multi-view fusion, depth data captured from multiple camera positions are merged into a unified volumetric or point-based representation. For example, Volumetric Fusion constructs a Truncated Signed Distance Function (TSDF) from sequential scans to recover occluded regions before grasp generation [22]. Active perception extends this concept by autonomously selecting camera viewpoints to reduce uncertainty. Reinforcement learning–based planners optimize view sequences for maximum information gain or grasp success, achieving up to 15-20% improvement over static scanning strategies [23]. Recent generative pipelines integrate both using uncertainty maps from diffusion or transformer based modules to guide active viewpoint planning and complete missing geometry before final grasp synthesis [7]. By combining perception and decision-making, these systems enhance robustness in cluttered environments and enable adaptive grasp planning in dynamic scenes. Table-2 shows Comparison of Occlusion Handling Architectures.

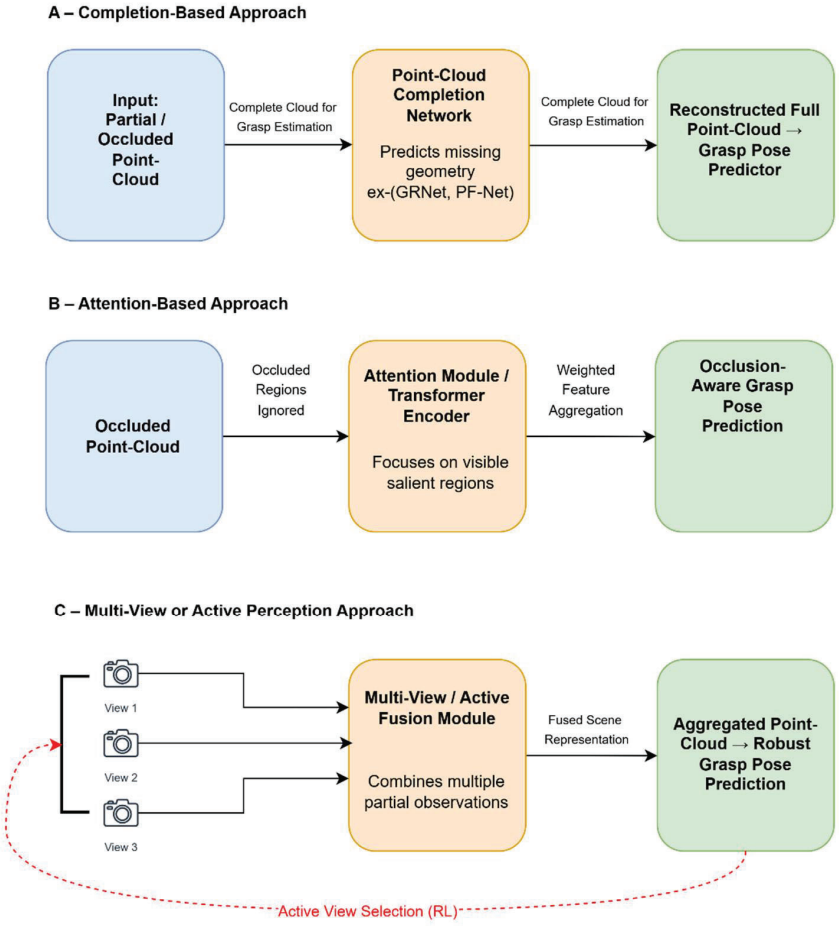

Fig.6. Occlusion Handling Strategies in Grasp Pose Detection

TABLE 2: COMPARISON OF OCCLUSION HANDLING ARCHITECTURES

Method	Input Modality	Planning Algorithm	Reported Grasp Success Improvement
Grasp-the-Graph	Scene point cloud	Graph-based next-best-view selection	+12% GSR under 60% occlusion
TransSC	Sequential partial point clouds	Spatiotemporal fusion + active NBV	+15% GSR, –25% reconstruction error
Contact-GraspNet	RGB-D + tactile contacts	Contact-aware grasp ranking	+18% GSR in cluttered scenes
Single-View Scene → Point Cloud	Single RGB-D frame	Hypothesis refinement via one additional view	+10% GSR with one extra view
DORA	RGB-D + affordance maps	RL-based active perception policy	+14% GSR under heavy occlusion

V. DATASETS AND EVALUATION

A. Benchmark Datasets

Large-scale datasets are fundamental to developing and benchmarking generative grasping models. They provide the necessary diversity, controlled environments, and annotations to support model training, evaluation, and reproducibility.

1) Synthetic Grasping Datasets

Synthetic datasets have accelerated progress in generative grasp synthesis by providing large volumes of structured grasp annotations that would be infeasible to collect physically. GraspNet-1Billion remains one of the most comprehensive, with over one billion 6-DOF grasps annotated across 88 object

categories and partial point clouds captured from multiple viewpoints for occlusion handling and completion analysis [23]. Dex-Net 2.0 provides millions of analytic grasps derived from force-closure models, enabling probabilistic grasp quality learning on depth inputs [10]. Datasets such as Jacquard and ShapeNet-Part serve as additional resources for planar and part-based grasp studies, while ACRONYM, TS-ACRONYM, and DexGraspNet extend the paradigm to dexterous, multi-finger grasping under simulated occlusion [7], [13]. Despite their advantages in scale and consistency, these synthetic datasets still exhibit a "reality gap" due to idealized physics and rendering environments. Table-3 provides a comparative summary of major synthetic datasets used in generative grasping research, highlighting their modality, scale, occlusion settings, and virtual environments. Table-3 shows Synthetic Datasets Comparison

TABLE 3: SYNTHETIC DATASETS COMPARISON

Dataset	Modality	#Objects	#Scenes/#Grasps/#Images	Supports Occlusion?	Sensor Types / Virtual Setup
GraspNet-1Billion	RGB-D, Point Cloud	88	97,280 images/1,000,000,000 grasps	Yes	Simulated Azure Kinect in Isaac Gym
ACRONYM	Point Cloud	8872	17.7 million grasps / 5,000 scenes	Yes	NVIDIA FleX simulation
DexGraspNet	Point Cloud	5355	1.32 million grasps / 8,270 scenes	Yes	Simulated Kinect v3 in NVIDIA Omniverse Robotics
GraspGen	Point Cloud	36000	53 million grasps	Yes	PyBullet with randomized lighting and textures
Dex-Net 2.0	Depth, Point Cloud	1000+	6.8 million grasps	Yes	Simulated Depth Cameras with analytic friction and mass parameters
TS-ACRONYM	RGB-D	88	97,280 images	Yes	Azure Kinect with synthetic noise injection
YCB-GraspNet	RGB-D	77	21,000 grasps	Yes	RealSense SR300 and Kinect v2, manual annotation
ShapeNetGrasp	Point Cloud	55000	2 million grasps	No	Mesh renderer with random camera views

2) Real-World Grasping Benchmarks

Real-world datasets evaluate the robustness of generative grasping systems under natural noise, illumination variation, and physical occlusion. The *YCB Object and Model Set* and *Cornell Grasp Dataset* remain standard references for assessing grasp detection and success under RGB-D sensing[23]. More complex benchmarks such as *YCB-Video* and *VMRD* introduce multi-object clutter and explicit occlusion relations [28], while *GraspNet-1Billion* extends its synthetic foundation with physical RGB-D captures from Intel RealSense and Azure Kinect cameras [29].As shown in Fig.7. While these datasets provide essential ground truth for real-robot experiments, they often lack the scale, annotation uniformity, and diversity available in simulation. Table-4 summarizes major real-world datasets, detailing sensing modalities, capture setups, and the extent of occlusion complexity.

B. Evaluation Metrics

A comprehensive evaluation of generative grasping models requires integrating geometric, physical, and temporal performance measures. The Grasp Success Rate (GSR) is the most widely used quantitative metric, representing the percentage of successful stable lifts in robotic or simulated executions [23]. Complementary metrics such as Force Closure (FC) and wrench-space stability analyses quantify mechanical robustness [4].For generative models that perform completion, Chamfer Distance and Earth Mover's Distance are employed to assess geometric fidelity between predicted and ground-truth shapes [11]. Inference time serves as a proxy for real-time feasibility frameworks such as *Heatmap-Guided 6-DOF Grasping* demonstrate sub-50 ms latency without major performance loss [30]. In occlusion-heavy settings, Visible Ratio and Reconstruction Error further capture accuracy under partial perception [31]. Together, these measures create a multi-dimensional view of generative model performance, balancing precision, robustness, and efficiency. Table-5 summarizes these metrics, their purposes, limitations, and typical ranges reported in current literature.

C. Protocols and Benchmark Practices

To ensure reproducibility and fair comparison, modern grasping research follows standardized evaluation protocols that bridge synthetic and real domains. In simulation-to-real transfer, models are pre-trained on synthetic datasets such as *Dex-Net* or *GraspNet-1Billion* and later fine-tuned or directly validated on physical captures using domain randomization or

TABLE 4: REAL-WORLD DATASETS COMPARISON

Dataset	Year	Modality (RGB / Depth / Point Cloud)	#Objects	#Scenes / #Images / #Grasps	Supports Occlusion?	Sensor Types / Capture Setup
Cornell Grasping Dataset	2011	RGB-D	240	1,035 images; 8,019 grasps	No (single-object scenes)	One Kinect camera; tabletop single-object captures
Pinto & Gupta 50K	2016	RGB-D	150	50,000 images; 50,000 grasps	No (single-object trials)	Single camera; robot trials for grasp attempts across multiple objects
YCB-Video (PoseCNN)	2017	RGB-D	21	92 scenes; ~134,000 frames	Yes (multi-object occlusion and clutter)	Single RGB-D camera; video sequences capturing multiple objects per scene
Levine et al. 800K (Hand-Eye Coordination)	2018	RGB-D	-	800,000 images; 800,000 grasps	No (mostly single-object in bin picking; minimal occlusion)	Single camera per robot; large-scale multi-robot data collection
VMRD (Visual Manipulation Relationship Dataset)	2018	RGB-D	— (multi-object scenes)	4,700 images; ~3 objects per scene	Yes (stacking/occlusion relations)	RGB camera; tabletop stacking scenes capturing manipulation relationships
Multimodal Grasp Dataset (RGB-D-T)	2019	RGB-D-T	single-object setting	2,550 samples	No (single-object scenes)	Controlled setup with RGB-D + thermal/tactile sensing
GraspNet-1Billion	2020	RGB-D, Point Cloud	88	190 scenes; 97,280 images; >1B grasps	Yes (multi-object clutter)	Multi-view capture (Intel RealSense D435 + Azure Kinect) on robot arm; 256 viewpoints/scene
SuctionNet-1Billion (suction grasping)	2021	RGB-D, Point Cloud labels	— (shared object set with GraspNet scenes)	97,000 images; 10 scenes per object; 3–8M grasps per scene	Yes (multi-object clutter)	Robot-mounted multi-view RGB-D capture; large-scale suction grasp annotations

noise augmentation to close the reality gap [13], [32]. Ablation studies help quantify the contribution of specific components, such as point-completion sub-networks or diffusion-based sampling modules, while cross-dataset testing measures adaptability when training on one dataset (e.g., *Dex-Net*) and evaluating on another (e.g., *YCB)* [4], [28]. Such practices foster transparent, consistent reporting and reveal how well generative grasping systems generalize to unseen conditions.

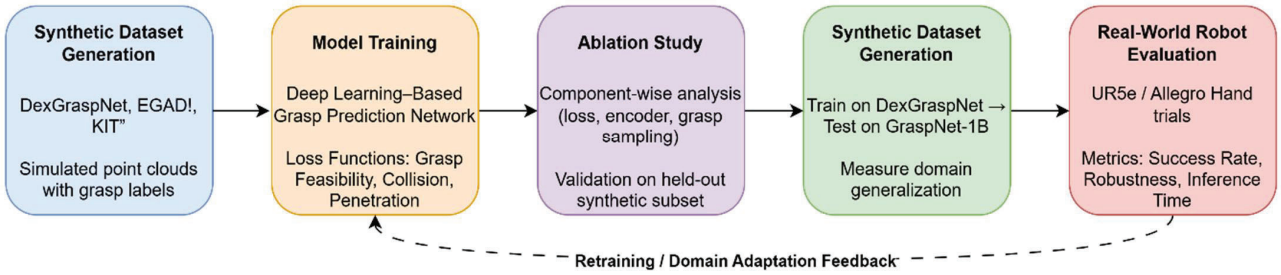

Fig.7. Dataset and Evaluation Flowchart for Grasp Pose Detection

TABLE 5: EVALUATION METRICS SUMMARY AND LIMITATIONS

Metric	Type	Definition	Use Case	Limitation	Typical Range / Example
Grasp Success Rate (GSR)	Execution	Percentage of executed grasps that successfully lift and hold the object	Overall system effectiveness	Varies with robot/controller; lacks standardization across studies	70–95% in simulation; 60-85% real robots
Force Closure (FC)	Execution	Fraction of grasps meeting analytic force-closure criterion	Stability assessment	Overestimates real-world success due to simplified friction/contact models	0–1 (binary / probability)
Average Precision (AP)	Prediction	Area under the precision–recall curve for predicted grasp candidates	Detection accuracy	Sensitive to score thresholds; does not reflect actual grasp execution	0.7–0.95 depending on dataset
Coverage	Prediction	Proportion of reachable object surface covered by grasp candidates	Grasp diversity analysis	High coverage may include low-quality or infeasible grasps	50–90%
Inference Time	Execution/Prediction	Average time between input capture and grasp output (ms)	Real-time feasibility	Hardware- and implementation-dependent; not directly comparable across studies	50–500 ms (GPU-based); >1s CPU-only
Collision Rate	Execution	Percentage of grasps causing unintended collisions before lift	Safety and reliability	Requires extensive simulation or real trials; underreported in many works	0–15% depending on clutter
Reconstruction Error	Prediction	Mean geometric error between predicted and ground-truth object or grasp shapes	Shape completion/ refinement quality	Dependent on ground-truth fidelity; not directly tied to grasp success	2-10mm (Chamfer Distance/ Hausdorff)

VI. COMPARATIVE PERFORMANCE ANALYSIS

Comparative evaluation of recent generative grasping frameworks reveals clear trade-offs among accuracy, generalization, and inference speed. Early latent-space models such as VAEs and GANs established the foundation for grasp sampling but struggled with precision and occlusion handling due to limited shape representation capabilities [3]. In contrast, diffusion-based methods (e.g., Grasp Diffusion Network, DexDiffuser) achieve higher robustness and grasp diversity by progressively refining samples through denoising steps. These methods consistently report grasp success rates above 90% in simulation benchmarks such as GraspNet-1Billion, though their iterative inference remains computationally heavy (50-200 ms per object) [3], [23].

TABLE 6(A): QUANTITATIVE COMPARISON OF PURELY GENERATIVE GRASPING METHODS

Method	Generative Model Type	Input Representation	Output Representation	Dataset Used	Grasp Success Rate / Accuracy (%)	Key Strengths	Key Limitations
DexDiffuser	Diffusion Model	Partial point cloud (Kinect v3)	16-DOF Allegro Hand grasp poses	DexGraspNet (1.7M grasps, 5,378 objects)	Sim: 98.77; Real: 68.89	High sampling diversity, end-to-end diffusion sampler + evaluator	Computationally heavy; sim-to-real gap
ZeroGrasp	Octree-based CVAE	Single RGB-D image	Joint 3D reconstruction + 6-DOF grasps	ZeroGrasp-11B (11.3B grasps, 12K objects) + GraspNet-1B	SOTA on GraspNet-1B	Zero-shot capability; handles transparent/specular surfaces	Compute-intensive; occlusion-sensitive
GraspGen	Diffusion-Transformer	Point cloud (object-centric)	6-DOF grasps for multiple grippers	GraspGen Dataset (53M grasps, 36K objects)	81.3 (real robot)	Multi-gripper compatibility; on-generator scoring filter	Long inference; gripper-specific
PSAGrasp	Transformer + Point Serialization	Serialized point cloud (Z-order)	6-DOF grasps with antipodal constraints	GraspNet-1Billion	Seen: 76.17; Similar: 70.59; Novel: 41.16	Attention-driven serialized points; antipodal grasp sampling	Serialization overhead; high transformer cost
Generalizing-Grasp	Domain Prior Knowledge + Diffusion	Multi-view TSDF	6-DOF grasps with constraints	GraspNet-1Billion	36.67 (novel)	Physically constrained diffusion; contact optimization	Multi-view requirement; slow
PhyGrasp	Multimodal Large Model (3D + Language)	Point cloud + natural language	Physics-informed grasp probability maps	PhyPartNet (195K part instances)	+10 over GraspNet baseline	Commonsense physical reasoning; language-conditioned	Dataset-limited; language dependency
Reasoning Grasping	Multimodal LLM + Vision	RGB + language	Instruction-conditioned grasp poses	Custom from GraspNet-1B	> baselines (CLIP/LLaVA) 120–180	Instruction-based grasp synthesis; multimodal reasoning	Early-stage; language-heavy

TABLE 6(B): QUANTITATIVE COMPARISON OF HYBRID GENERATIVE GRASPING METHODS

Method	Generative Model Type	Input Representation	Output Representation	Dataset Used	Grasp Success Rate / Accuracy (%)	Inference Time (ms)	Key Strengths	Key Limitations
Contact-GraspNet	Direct Regression (4-DOF reduced)	Depth / Point cloud	6-DOF grasps via contact points	ACRONYM (17M grasps, 8,872 objects)	> 90 (cluttered scenes)	190–280	Collision-aware; contact-driven; end-to-end	Limited to parallel-jaw; occlusion challenges
GraspNet-1Billion	Analytical + CNN prediction	Point cloud (RGB-D)	6-DOF grasps with approach vectors	GraspNet-1Billion (>1B grasps, 88 objects)	Benchmark baseline	Not specified	Large-scale, unified evaluation benchmark	Not generative; limited object diversity
FlexLoG	Region-based Local Grasp Model	RGB-D + local/global cues	Regional 6-DOF grasps	GraspNet-1Billion (regional subset)	95 (real robot)	~38 (26 FPS)	Real-time; grasp-centric; flexible guidance	Requires explicit guidance input
GSNet	Graspness-based Generation	Point cloud	Graspness score + 6-DOF poses	GraspNet-1Billion	Seen: 67.12; Similar: 54.81; Novel: 24.31	Real-time	Simple and efficient graspness concept	Weak performance on novel objects
HGGD	Heatmap-Guided CNN	RGB-D + heatmap	6-DOF grasps	GraspNet-1Billion + TS-ACRONYM	Seen: 59.36; Similar: 51.20; Novel: 22.17	~36 (28 FPS)	Fast; local-global semantic fusion	Heatmap dependency; lower novel accuracy

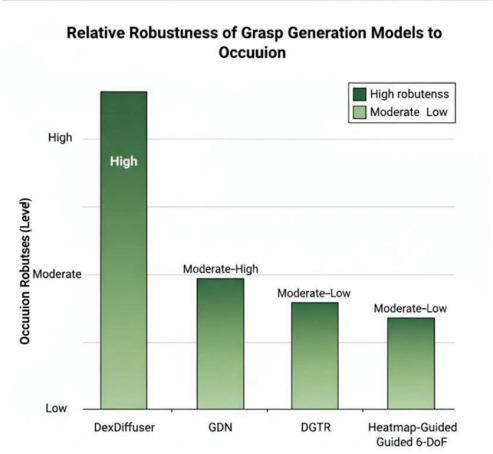

Fig.8. Performance vs. Occlusion Level

Diffusion-based frameworks such as *Grasp Diffusion Network* yield diverse and robust grasps under occlusion but remain computationally demanding [6], [7]. Transformer-based models, including *Dexterous Grasp Transformer (DGTR)*, offer a practical compromise between accuracy and speed through efficient attention mechanisms [34]. Regression and heatmap-guided networks achieve real-time performance yet lack grasp diversity and stability in cluttered or partially observed scenes [30]. Synthetic-data-trained systems require domain adaptation for real-world consistency [23], while multimodal designs like *Transformer-Tactile Robot Grasp* enhance adaptability at the cost of sensing and calibration complexity [35].Fig.8 shows graph of Performance vs. Occlusion Level.

VII. TECHNICAL INSIGHTS

Generative models for grasp synthesis have evolved around a set of core design philosophies that balance accuracy, diversity, and computational efficiency. While architectures vary across studies, most share a common objective to generate physically feasible and context-aware grasp configurations even when data is incomplete or uncertain.

A. Architectural Design Patterns

Recent models reveal clear trends in architectural design. A first principle is latent feature fusion, where global geometric cues and local surface details are combined within a shared representation. This strategy allows networks to reason jointly about overall object shape and contact-level constraints, as demonstrated in *Constrained 6-DoF Grasp Generation* [21]. A second pattern, hierarchical encoding, organizes feature extraction at multiple resolutions coarse geometry, mid-level part structure, and fine contact regions. This structure, employed in diffusion-based models such as *DexDiffuser* [22], supports both global grasp reasoning and local precision.

Finally, multi-branch processing has become increasingly common, particularly in multimodal grasp frameworks that incorporate RGB, depth, or tactile inputs through specialized sub-networks before feature fusion [35]. These architectural choices collectively improve robustness under occlusion and variability, establishing a foundation for scalable and adaptive generative grasping frameworks.

B. Loss and Training Strategies

Training procedures for generative grasp models combine physical, geometric, and perceptual constraints. A core component is the grasp feasibility loss, which enforces contact stability using force-closure or wrench-based conditions [21]. This ensures that generated grasps remain physically realizable. In addition, multi-task learning links grasp generation with auxiliary objectives such as object completion or affordance prediction, improving feature generalization under sparse data conditions [37]. Models are often refined using semi-supervised or self-supervised consistency losses, where unlabelled data contributes through reconstruction or contrastive objectives [38]. Bridging the gap between simulation and real-world deployment remains a major concern. Approaches such as domain randomization**,** sensor noise injection**,** and progressive adaptation have proven effective in improving real-world transferability [14]. Such strategies align learning objectives with both physical realism and practical robustness, helping generative models achieve reliable real-world grasp performance. Representative Architecture Flow Diagram shown in Fig.9.

C. Computational and Deployment Aspects

Practical deployment requires balancing grasp quality with computational cost. Diffusion-based frameworks provide the highest grasp diversity but remain computationally intensive, often requiring several hundred milliseconds per sample [13]. Transformer-based approaches, such as *Dexterous Grasp Transformer*, offer a middle ground with inference times near 30–80 ms while maintaining accuracy [34]. Heatmap-guided and regression-based models remain the fastest, achieving sub-50 ms predictions suitable for real-time control [30]. Hardware constraints further influence design. Lightweight encoders, quantization-aware training, and pruning are increasingly adopted to enable real-time inference on embedded systems without significant loss in grasp reliability [23]. Table 7 summarizes the computational metrics average inference time, peak memory usage, and model parameter count of leading generative methods, providing guidance for robotics practitioners balancing performance and resource constraints. The trend toward compact and adaptive architectures reflects a broader movement in robotics research: enabling high-level reasoning on limited hardware, paving the way for practical deployment of generative grasping in autonomous systems.

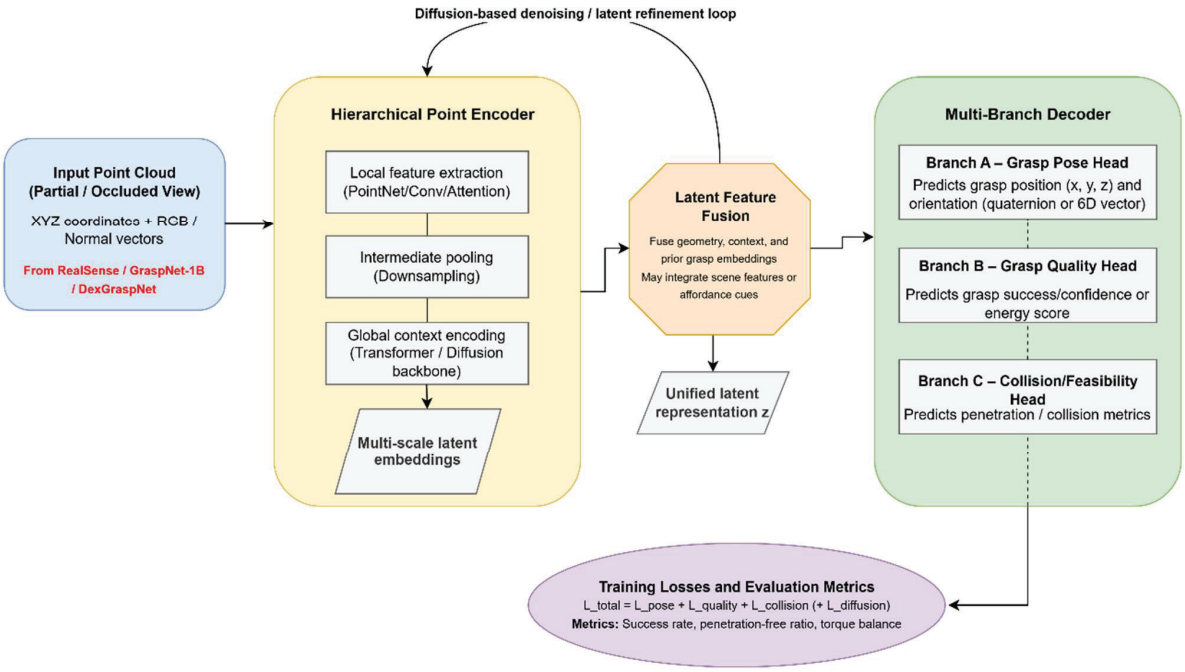

Fig.9. Representative Architecture Flow Diagram Illustrating Latent Fusion, Hierarchical Encoding, and Multi-Branch Design

TABLE 7: COMPUTATIONAL EFFICIENCY COMPARISON

Model	Inference Time (ms)	Peak Memory Usage (MB)	Model Parameters (Millions)	Edge Deployment Feasibility
DexDiffuser	120–180	~4500	~85	Low (GPU recommended)
Grasp Diffusion Network	150–250	~3800	~72	Low (GPU recommended)
GraspGen	2500–3000	~6200	~120	Very Low (workstation-class GPU)
Dexterous Grasp Transformer, DGTR	50–80	~2100	~45	Moderate (edge GPU feasible)
Transformer-Grasp, TF-Grasp	30–60	~1800	~38	High
Contact-GraspNet	190–280	~1500	~28	High
HGGD, Efficient Heatmap-Guided 6-DoF	28–36 (≈28 FPS)	~900	~15	Very High

VIII. APPLICATIONS AND USE CASES

Generative grasping frameworks have shown broad applicability across industrial, service, and healthcare domains, primarily due to their ability to reason under uncertainty and adapt to partial or noisy observations. In industrial and service settings, diffusion- and transformer-based grasp generators are used for bin-picking, sorting, and assembly, where objects vary in shape and orientation. These models provide multiple feasible grasp hypotheses, improving task success rates without manual reprogramming [23], [35]. By integrating vision and tactile feedback, robots achieve higher dexterity in handling irregular or fragile items, which is crucial for both warehouse automation and assistive tasks. In healthcare and assistive robotics, similar frameworks support safe interaction with deformable or delicate objects such as medical tools, packaging, or food items. Their capacity to model uncertainty and plan adaptive grasps ensures reliability in unstructured human environments [6], [23].

Simulation platforms such as Isaac Gym, PyBullet, and NVIDIA Omniverse continue to accelerate research on these systems, allowing scalable evaluation of model generalization and transferability [13], [28]. Standardized datasets like GraspNet-1Billion and DexGraspNet further promote benchmarking and reproducibility across generative architectures [10].

IX. CHALLENGES AND OPEN PROBLEMS

Despite notable advancements, several challenges continue to restrict the broader deployment of generative grasping systems. Robustness under occlusion remains a primary concern performance drops significantly in cluttered or partially observed scenes, emphasizing the need for improved multi-view fusion, shape completion, and uncertainty-aware inference [23]. Another persistent challenge is cross-domain generalization. Models trained primarily on synthetic datasets often fail to adapt to real-world variability in object texture, lighting, and material properties unless extensive domain adaptation or fine-tuning is applied [36]. Training data scarcity further limits progress. The collection of large, annotated grasp datasets that reflect realistic diversity is still resource-intensive, slowing the development of generalizable frameworks [37]. Finally, computational efficiency remains an obstacle to real-time robotic deployment. High-performing generative models especially diffusion and transformer-based demand heavy computation, creating a trade-off between accuracy and latency [13].

X. FUTURE RESEARCH DIRECTIONS

Future work in generative grasping focuses on enhancing generalization, adaptability, and real-time performance. The adoption of large pre-trained foundation models offers prospects for zero-shot grasp synthesis and improved transferability across unseen objects. Vision language models, such as CLIP-Grasp, bring semantic reasoning into manipulation, enabling task-aware grasping beyond geometric matching [38]. Advances in self-supervised learning and hardware-aware optimization are expected to improve efficiency on embedded robotic systems [39]. Equally important is the need for standardized benchmarks and reproducibility protocols to ensure consistent comparison and accelerate progress across studies [40].

XI. CONCLUSION

Generative deep learning has brought a significant shift in robotic grasp pose detection, allowing models to learn diverse and physically consistent grasp strategies directly from point cloud data. Recent frameworks based on diffusion models and transformers have demonstrated impressive improvements in grasp diversity, stability, and adaptability, especially under partial or occluded observations. These architectures integrate hierarchical encoding, latent fusion, and multi-branch prediction to achieve higher precision and robustness across varied object types and environments. Despite these advancements, several persistent challenges remain. Grasp success often degrades under severe occlusion or clutter, while domain generalization issues limit transfer from synthetic to real-world settings. The scarcity of labelled real-world data and the high computational demands of generative networks further constrain scalability and real-time application.

Ongoing research is progressively addressing these gaps through multi-view fusion, domain adaptation, self-supervised training, and model optimization strategies. Building richer and more standardized datasets, along with developing efficient architectures suitable for embedded systems, will be key to enabling practical deployment. As these directions converge, generative grasping frameworks are poised to evolve into robust, adaptive, and efficient systems bridging the gap between simulation and reality, and enabling robots to manipulate their environments with human-like precision and resilience.

REFERENCES

[1] H. Zhang, J. Tang, S. Sun, and X. Lan, “Robotic Grasping from Classical to Modern: A Survey,” Feb. 2022, [Online]. Available: http://arxiv.org/abs/2202.03631

[2] Z. Xie, X. Liang, and C. Roberto, “Learning-based robotic grasping: A review,” 2023, *Frontiers Media S.A.* doi: 10.3389/frobt.2023.1038658.

[3] K. Kleeberger, R. Bormann, W. Kraus, and M. F. Huber, “A Survey on Learning-Based Robotic Grasping,” *Current Robotics Reports*, vol. 1, no. 4, pp. 239–249, Dec. 2020, doi: 10.1007/s43154-020-00021-6.

[4] R. Newbury *et al.*, “Deep Learning Approaches to Grasp Synthesis: A Review,” May 2023, [Online]. Available: http://arxiv.org/abs/2207.02556

[5] Y. Guo, H. Wang, Q. Hu, H. Liu, L. Liu, and M. Bennamoun, “Deep Learning for 3D Point Clouds: A Survey,” Jun. 2020, [Online]. Available: http://arxiv.org/abs/1912.12033

[6] J. Carvalho *et al.*, “Grasp Diffusion Network: Learning Grasp Generators from Partial Point Clouds with Diffusion Models in SO(3)xR3,” Dec. 2024, [Online]. Available: http://arxiv.org/abs/2412.08398

[7] R. Wolf, Y. Shi, S. Liu, and R. Rayyes, “Diffusion models for robotic manipulation: a survey,” 2025, *Frontiers Media SA*. doi: 10.3389/frobt.2025.1606247.

[8] H. Gui *et al.*, “High–performance grasp pose detection via point cloud serialization attention,” *Pattern Recognit*, vol. 171, Mar. 2026, doi: 10.1016/j.patcog.2025.112099.

[9] M. Q. Mohammed, K. L. Chung, and C. S. Chyi, “Review of deep reinforcement learning-based object grasping: Techniques, open challenges, and

recommendations," *IEEE Access*, vol. 8, pp. 178450–178481, 2020, doi: 10.1109/ACCESS.2020.3027923.

[10] P. Xie, S. Chen, W. Tang, D. Hu, W. Yang, and G. Wang, "Rethinking 6-Dof Grasp Detection: A Flexible Framework for High-Quality Grasping," Oct. 2024, [Online]. Available: http://arxiv.org/abs/2403.15054

[11] W. Yuan, T. Khot, D. Held, C. Mertz, and M. Hebert, "PCN: Point Completion Network," Sep. 2019, [Online]. Available: http://arxiv.org/abs/1808.00671

[12] Z. Weng, H. Lu, D. Kragic, and J. Lundell, "DexDiffuser: Generating Dexterous Grasps with Diffusion Models," Nov. 2024, [Online]. Available: http://arxiv.org/abs/2402.02989

[13] W. Zhao, J. P. Queralta, and T. Westerlund, "Sim-to-Real Transfer in Deep Reinforcement Learning for Robotics: a Survey," Jul. 2021, doi: 10.1109/SSCI47803.2020.9308468.

[14] H. Wang, Y. Zhang, Y. Wang, and J. Li, "6-DoF Grasp Detection in Clutter with Enhanced Receptive Field and Graspable Balance Sampling," Dec. 2024, [Online]. Available: http://arxiv.org/abs/2407.01209

[15] X. Song, Y. Li, Y. Zhang, Y. Liu, and L. Jiang, "An overview of learning-based dexterous grasping: recent advances and future directions," *Artif Intell Rev*, vol. 58, no. 10, Oct. 2025, doi: 10.1007/s10462-025-11262-2.

[16] A. Murali *et al.*, "GraspGen: A Diffusion-based Framework for 6-DOF Grasping with On-Generator Training," Jul. 2025, [Online]. Available: http://arxiv.org/abs/2507.13097

[17] Z. Ding, Y. Sun, S. Xu, Y. Pan, Y. Peng, and Z. Mao, "Recent Advances and Perspectives in Deep Learning Techniques for 3D Point Cloud Data Processing," Aug. 01, 2023, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: 10.3390/robotics12040100.

[18] H. Ma, M. Shi, B. Gao, and D. Huang, "Generalizing 6-DoF Grasp Detection via Domain Prior Knowledge." [Online]. Available: https://github.com/mahaoxiang822/

[19] Y. Zhou *et al.*, "EAGA-Net: a novel simulation-based grasping detection dataset and network with efficient adaptability of gripper attribute," *Advanced Engineering Informatics*, vol. 68, Nov. 2025, doi: 10.1016/j.aei.2025.103600.

[20] H. Gui *et al.*, "High–performance grasp pose detection via point cloud serialization attention," *Pattern Recognit*, vol. 171, Mar. 2026, doi: 10.1016/j.patcog.2025.112099.

[21] G. Singh *et al.*, "Constrained 6-DoF Grasp Generation on Complex Shapes for Improved Dual-Arm Manipulation," Jul. 2024, [Online]. Available: http://arxiv.org/abs/2404.04643

[22] T.-T. Le, T.-S. Le, Y.-R. Chen, J. Vidal, and C.-Y. Lin, "6D Pose Estimation with Combined Deep Learning and 3D Vision Techniques for a Fast and Accurate Object Grasping."

[23] H.-S. Fang, C. Wang, and M. G. C. Lu, "GraspNet-1Billion: A Large-Scale Benchmark for General Object Grasping." [Online]. Available: www.graspnet.net.

[24] W. Chen, H. Liang, Z. Chen, F. Sun, and J. Zhang, "Improving Object Grasp Performance via Transformer-Based Sparse Shape Completion," *Journal of Intelligent and Robotic Systems: Theory and Applications*, vol. 104, no. 3, Mar. 2022, doi: 10.1007/s10846-022-01586-4.

[25] M. Sundermeyer, A. Mousavian, R. Triebel, and D. Fox, "Contact-GraspNet: Efficient 6-DoF Grasp Generation in Cluttered Scenes," Mar. 2021, [Online]. Available: http://arxiv.org/abs/2103.14127

[26] Y.-K. Wang, C. Xing, Y.-L. Wei, X.-M. Wu, and W.-S. Zheng, "Single-View Scene Point Cloud Human Grasp Generation." [Online]. Available: https://github.com/iSEE-Laboratory/S2HGrasp.

[27] L. Zhang *et al.*, "DORA: Object Affordance-Guided Reinforcement Learning for Dexterous Robotic Manipulation," May 2025, [Online]. Available: http://arxiv.org/abs/2505.14819

[28] G. Du, K. Wang, S. Lian, and K. Zhao, "Vision-based Robotic Grasping From Object Localization, Object Pose Estimation to Grasp Estimation for Parallel Grippers: A Review," Oct. 2020, doi: 10.1007/s10462-020-09888-5.

[29] A. K. Sharma and J. Kunkel, "Grasp Pose Detection in Point Clouds," May 2025, doi: 10.1177/ToBeAssigned.

[30] S. Chen, W. Tang, P. Xie, W. Yang, and G. Wang, "Efficient Heatmap-Guided 6-Dof Grasp Detection in Cluttered Scenes," May 2024, [Online]. Available: http://arxiv.org/abs/2403.18546

[31] Z. Yin and Y. Li, "Overview of robotic grasp detection from 2D to 3D," Jan. 01, 2022, *KeAi Communications Co.* doi: 10.1016/j.cogr.2022.03.002.

[32] W. Zhu, X. Guo, D. Owaki, K. Kutsuzawa, and M. Hayashibe, "A Survey of Sim-to-Real Transfer Techniques Applied to Reinforcement Learning for Bioinspired Robots," *IEEE Trans Neural Netw Learn Syst*, vol. 34, no. 7, pp. 3444–3459, Jul. 2023, doi: 10.1109/TNNLS.2021.3112718.

[33] C. Guo *et al.*, "End-to-End lightweight Transformer-Based neural network for grasp detection towards fruit robotic handling," *Comput Electron Agric*, vol. 221, Jun. 2024, doi: 10.1016/j.compag.2024.109014.

[34] G.-H. Xu, Y.-L. Wei, D. Zheng, X.-M. Wu, and W.-S. Zheng, "Dexterous Grasp Transformer." [Online]. Available: https://github.com/iSEE-Laboratory/DGTR.

[35] E. Y. Puang, Z. Li, C. M. Chew, S. Luo, and Y. Wu, "Learning Stable Robot Grasping with Transformer-based Tactile Control Policies," Jul. 2024, [Online]. Available: http://arxiv.org/abs/2407.21172

[36] J. Bohg, A. Morales, T. Asfour, and D. Kragic, "Data-driven grasp synthesis-A survey," *IEEE Transactions on Robotics*, vol. 30, no. 2, pp. 289–309, 2014, doi: 10.1109/TRO.2013.2289018.

[37] P. Xu and Y. Mu, "Weakly-Supervised Affordance Grounding Guided by Part-Level Semantic Priors," May 2025, [Online]. Available: http://arxiv.org/abs/2505.24103

[38] S. Xia, D. Chen, R. Wang, J. Li, and X. Zhang, "Geometric Primitives in LiDAR Point Clouds: A Review," 2020, *Institute of Electrical and Electronics Engineers*. doi: 10.1109/JSTARS.2020.2969119.

[39] S. Jin, J. Xu, Y. Lei, and L. Zhang, "Reasoning Grasping via Multimodal Large Language Model," Oct. 2024, [Online]. Available: http://arxiv.org/abs/2402.06798

[40] "The Comprehensive Review of Vision-based Grasp Estimation and Challenges."