RGB-D Bin-Picking System for Ergonomic Automotive Clip Assembly: 3D Annotation, Deep Learning, and 6-DOF Pose

Brahim Bergor Beguiel[1], Ibrahim Hadj Baraka[2], Yassir Zardoua[3]
Department of Electrical Engineering-Faculty of Science and Techniques,
Abdelmalek Essaâdi University, Morocco[1,2]
3000 Avenue Saint Louis, 33000 Fès, Morocco[3]

***Abstract*—Automotive seat cover assembly entails the manual collection of plastic J-shaped clips from boxes located at a non-ergonomic distance from the sewing stations. Operators are compelled to inadvertently pick up multiple components at once, and make actions that involve repetitive stretching. The daily, consistent repetition leads to misassembly of the critical seat to frame connectors, and adds on to physical stress. This study showcases a fully integrated RGB-D bin-picking solution that uses depth-dependent grasp planning and deep learning object detection for interlocked plastic clip handling. GraspAnnotator Pro, a dual-modality bespoke software solution developed for this work, allows for model training on point cloud and RGB data for cluttered environment 6D pose estimation. This is achieved through a custom integrated annotation tool that simplifies the labeling of grasp pose and object boundary assignments. The system reduces strain on operators by automatically positioning parts within ergonomic zones and using fault-tolerant handling integrated with assembly verification. Real-world deployment validation over 6 weeks of continuous operation accumulated 3,000 pick-and-place cycles across 10 distinct J-shaped wire harness components, achieving a 93.7% first-attempt success rate with an average cycle time of 10.6 seconds. The system demonstrates a 42% reduction in cycle time compared to manual methods (18.3 seconds) with significant ergonomic improvements.**

***Keywords*—Automated pick-and-place; RGB-D vision; YOLOv8 segmentation; point cloud matching; 6-DOF pose estimation; industrial robotics; automotive manufacturing; deep learning; quality inspection; PLC control**

I. Introduction

The automotive industry is undergoing a fundamental transformation driven by Industry 4.0 principles, where cyber-physical systems, intelligent automation, and real-time adaptive decision-making are reshaping traditional manufacturing processes. This paradigm shift is particularly relevant in labor-intensive assembly operations, where the integration of robotic systems with advanced perception capabilities and deep learning algorithms offers significant potential for addressing longstanding ergonomic and quality challenges.

Seat cover assembly is a labor intensive task, also in the automotive field, and must be done accurately and reliably to ensure that structure remains structurally sound for passenger protection. The process also involves "J" shaped extruded plastic pieces, as in Fig. 1, that work much like mechanical fasteners and retain fabric covers onto metal seat frames for the life of the vehicle.

Fig. 1. Automotive seat cover showing J-shaped plastic retainers at different attachment points (P1-P4) that secure fabric to metal frame.

Such accessories, made on a continuous extrusion equipment, exhibit inherent dimensional variations created by material shrinkage, temperature changes and wear of the extrusion die. Due to the interconnected structure of such members, produced for storage compactness, substantial handling complexity is introduced by this configuration when in stored form en masse in industrial bins. Today the majority of manufacturers depend on manual assembly to assemble J-clips, where an operator reaches into separate bins to pick up individual ones and these bins are placed at non-ergonomic distances from the sewing machine workstations. The current manual assembly workstation is illustrated in Fig. 2. The layout of this configuration is such that it will compel an operator to repeatedly reach across a wide restricted area laden

Fig. 2. Current manual assembly workstation showing sewing machine with non-ergonomic component placement requiring repetitive operator reaching.

with fabric material far from conventional ergonomic zones as defined by ISO 11228-3:2007, causing muscle fatigue and lost productivity.

More importantly, the interlocking nature of stored parts often causes unintentional multi-part grasping where operators retrieve 1–3 parts at a time. This issue leads to errors on part mis-identification, assembly steps deviation and parts missing—each of these being critical quality issues because of the load-bearing role that these components play in seat cover-to-frame attach systems. In addition, high-mix, low-volume production mode can be observed in automotive seat manufacturing where one has more than one component variation for each vehicle model and that multi-variant configurations between components of different shapes are introduced to pose a heavier cognitive burden on operators when selecting parts and verifying the assembly quality.

Robotic bin-picking is a promising technology for handling these challenges and extensive work has been conducted to prove the effectiveness of RGB-D vision systems in combination with deep learning algorithms towards object detection and pose estimation [1], [2], [3]. Modern methods use convolutional neural networks (CNNs) or transformer-based models as the most effective technique for 2D object localization [19], and recent research have proposed data-driven grasping based on synthetic training with various object geometries. Although prior bin-picking work has mainly considered applications where the objects to be picked are isolated and well separated, with relatively consistent geometrical form [9], [10], [8], the special cases of interlocking J-clips with profile overlap, variety in internal locking methods and inter-material occlusions have not been covered widely in literature. This work addresses this gap by presenting an integrated bin-picking system specifically designed for interlocked, dimensionally-variant flexible components.

This study addresses these gaps through three principal contributions. First, we introduce GraspAnnotator Pro, a novel dual-modality annotation tool that synchronously extracts RGB polygon annotations for YOLOv11-based object detection [20] alongside depth information from point cloud data to determine optimal suction cup contact positions. This annotation framework explicitly addresses the interlocked bin-picking scenario by enabling training datasets that capture the geometric complexity and pose variability inherent to interlocked J-clips. Second, our contributions include a vision guided bin-picking system which coordinates with a sewing machine. The system involves a pre-assembly inventory scan for efficient picking order and a fault-tolerant system that immediately swaps out defective components if quality control indicates a defect. Plucked accessories are within ISO 11228-3:2007 ergonomic reach bands of the operators in order to avoid unnecessary operator movements and provide consistent accessibility of components to the operators. Third, we offer comprehensive experimental validation through 6-week production deployment accumulating 3,000 cycles with 93.7% first-attempt success rate, 42% cycle time reduction, and 62% ergonomic improvement compared to manual assembly.

II. Literature Review

The automation of automotive assembly operations increasingly focuses on robotic bin-picking systems for handling small fastening components that present unique challenges due to dimensional variability and physical interlocking. This review examines current research in RGB-D vision-based manipulation, dataset creation and annotation workflows, ergonomic considerations, multi-modal sensor fusion, and fault-tolerant control, systematically identifying critical gaps motivating specialized bin-picking solutions for automotive seat assembly.

A. RGB-D Vision and 6D Pose Estimation

Several state-of-the-art methods have been developed for cluttered industrial scenes in recent years. Van Nguyen et al. captured real-world industrial clutter, including occlusions, challenging geometries, and reflective surfaces, to create the novel XYZ-IBD dataset, which they used to achieve state-of-the-art performance using symmetry-aware evaluation metrics [1]. Li et al. leveraged sim-to-real techniques to develop methods for industrial bin-picking tasks trainable through iterative self-training [2]. Kleeberger et al. developed a real-time single-shot 6D pose estimation method [3]. Point cloud-based techniques have also been developed to overcome the limitations of RGB information for industrial objects that can have surface rust, color variation, and lack of texture.

Recent general-purpose grasping frameworks have further advanced the field. Contact-GraspNet [16] generates grasp poses directly from depth point clouds using contact-based grasp representation, achieving robust performance on diverse object categories. AnyGrasp [17] proposes a large-scale foundation model for grasp detection that demonstrates strong generalization across geometries by learning from large synthetic datasets with real-world fine-tuning. FoundationPose [18] represents a significant advance in model-free 6D pose estimation, leveraging foundation model representations to generalize to novel objects without object-specific training [21]. These methods, however, are evaluated primarily on isolated objects with distinct geometric primitives and do not explicitly address the challenges of mechanically interlocked components with dimensional variability inherent to automotive fasteners.

Few methods consider that objects may have different geometry from their CAD models, and there are often tight tolerances in object dimensions ($\pm$0.5-2mm for fasteners used

in the automotive industry). Dimensional variations are often not explicitly modeled, but are rather treated as noise that needs to be robust to. In addition, most methods do not address scenarios with physically interlocked objects where multiple components remain mechanically connected.

B. Dataset Creation and Annotation Workflows

Deep learning approaches for robotic grasping critically depend on large-scale, precisely annotated datasets combining RGB appearance with accurate 3D geometric annotations [7], [8], [22]. However, creating datasets for non-canonical geometries encounters fundamental bottlenecks. Traditional annotation workflows universally assume pre-calibrated camera systems and rely on 2D image-based polygon selection, proving inadequate for capturing spatial complexity of curved geometries and identifying optimal contact surfaces.

Analysis of existing grasping datasets reveals significant gaps in coverage of non-standard geometries. Objects that are J-shaped, including hooks, carabiners, and wireframe fasteners, represent challenging cases for automated grasping but remain underrepresented in training data. Examination of major public datasets shows limited representation of such geometries, with most datasets focusing on rigid objects with well-defined geometric primitives. Furthermore, these datasets typically do not contain segment-level grasp annotations or suction zone specifications necessary for pneumatic grasping systems.

The annotation challenge is compounded by the need for precise 3D spatial information. Standard 2D annotation tools like LabelMe [4], CVAT [15], and commercial platforms such as Labelbox [23] require manual alignment between RGB images and depth data, introducing registration errors that degrade annotation quality. Recent work [4], [14] has explored integrated annotation frameworks that address these challenges through unified 3D annotation spaces, though the integration of dimensional variability within training pipelines remains an open challenge for ensuring robust generalization across different part sizes and configurations.

C. Research Gaps and Contributions

This review identifies six critical gaps representing significant barriers to effective small-component assembly automation:

- Dimensional Variability as Design Consideration: Existing systems assume geometrically consistent objects, treating 5-10% manufacturing dimensional variability as noise rather than intrinsic characteristics requiring explicit handling.
- High-Quality Training Data: Systematic integration of segment-specific grasp annotations and metric suction zone specifications into deep learning training pipelines remains limited.
- Mechanical Interlocking: Current research addresses visual occlusion but inadequately handles mechanical interlocking requiring sequential manipulation and force-based separation.
- Ergonomics as Primary Design Driver: Few systems architect ergonomic optimization as a primary objective driving operational strategies.
- Unified Dual-Modality Training: Specialized frameworks jointly optimizing RGB and point cloud modalities for precise 6D pose estimation in cluttered industrial scenarios remain limited.
- Integrated Fault-Tolerant Assembly Pipeline: Existing research addresses subsystem failures in isolation, lacking holistic fault tolerance spanning complete bin-picking-to-assembly workflows.

III. Custom Annotation Software and Deep Learning Framework

This section presents our comprehensive annotation and training framework, including: 1) a custom RGB-D annotation tool with pneumatic grasping circle support operating on colored point clouds, 2) calibrated projection system for precise RGB-to-3D alignment, 3) multi-modal deep learning architecture for 6-DOF pose estimation with adaptive suction radius prediction, and 4) complete training pipeline with deployment integration.

A. GraspAnnotator Pro: Advanced RGB-D Annotation Software

1) Motivation and design principles: The existing tools for annotating, such as LabelMe [4] and CVAT [15], primarily target 2D image annotations or basic 3D object detection. They lack specific functionality in handling the task of grasping in an industrial setting with pneumatic suction cups. Tools like PointNet++ [6] provide 3D feature learning capabilities but do not offer integrated annotation frameworks for grasp planning. Similarly, while Dex-Net [7] demonstrates deep learning-based grasp planning with synthetic point clouds, it does not address the challenge of annotating real-world RGB-D data for suction-based grasping in cluttered industrial scenarios. The project at hand requires: a single interface allowing the processing of both 3D and RGB data simultaneously, a method to define the grasping circle with the suction cup at an optimal position and radius in 3D space, automatic assignment of an RGB image to its point cloud through calibrated camera parameters [5], accurate mapping of RGB images to 3D through calibrated perspective projection, and a robust interface designed specifically for an industrial environment.

Building on these requirements, the latest version of GraspAnnotator Pro introduces a paradigm shift through a new way of doing polygon annotations, where polygons are directly annotated on a calibrated colored point cloud, not on an RGB image. All the problems associated with aligning 2D and 3D data at the same time are bypassed with this approach. Annotators can easily sketch polygons on the 3D representation of objects because the object shape and appearance can be viewed simultaneously within the unified spatial framework.

2) Software architecture and implementation: GraspAnnotator Pro, in Fig. 3, is developed with Python 3.9 and the graphical interface is based on Tkinter. The major libraries used are Open3D for point cloud processing and visualization, PIL/Pillow for image processing, NumPy for vector computations, and SciPy for geometric processing. To support the unified 3D annotation workflow, the interface is structured with three panes to maximize the efficiency of the annotation process: the left panel occupying 40% for the colored point

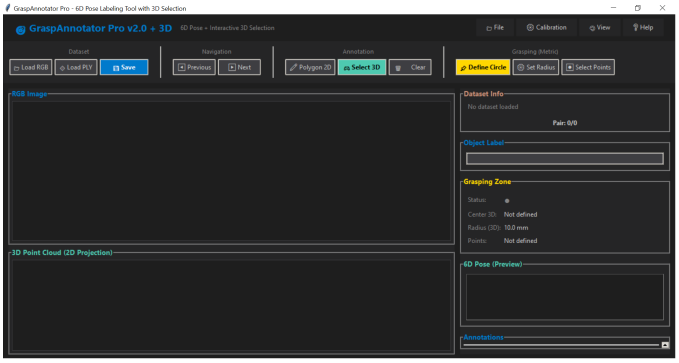

Fig. 3. GraspAnnotator pro interface showing colored point cloud view (left), annotation tools (middle), and data display (right).

cloud view to prioritize spatial visualization, the middle panel occupying 35% for the interactive annotation tools including polygon drawing and suction zone definition, and the right panel occupying 25% for the data displays.

a) Annotation workflow: The annotation process follows a structured five-step pipeline designed to minimize user effort while maximizing 3D spatial accuracy. *Step 1 — File Loading:* the operator loads a paired RGB-PLY scene; the automatic file-pairing module (Section III-A3) resolves file associations without user intervention. *Step 2 — Point Cloud Colorization:* the system projects RGB color onto the point cloud using the calibrated intrinsic and extrinsic parameters, producing a unified colored 3D view rendered as an orthographic projection. *Step 3 — Polygon Segmentation:* the annotator draws a closed polygon directly on the colored point cloud view; the tool resolves 3D coordinates for each vertex via k-d tree lookup and highlights the enclosed region with a semi-transparent overlay. *Step 4 — Suction Zone Definition:* the annotator clicks the optimal suction cup contact point within the segmented region and adjusts the radius via an interactive slider; the tool previews the selected circular contact zone in gold and computes the surface normal via PCA on the k-nearest neighbors. *Step 5 — Export:* the annotation is serialized to JSON containing 3D polygon vertices, suction circle center (x, y, z), radius in millimeters, surface normal vector (n_x, n_y, n_z), and point indices, ready for direct ingestion into the GraspingNet training pipeline. A session manager allows pausing and resuming annotation across multiple sessions, and batch export generates dataset manifests compatible with standard deep learning frameworks.

Calibrated three-dimensional point clouds are shown in the colored point cloud view as two-dimensional orthographic projections to preserve geometric measurements with the entire range of RGB data. Annotating is done directly on this view by drawing polygons, which gives the annotators immediate feedback in the form of semi-transparent overlays and vertices. The system also allows for the selection of points for grasping circles by means of radius sliders with real-time 3D preview. There is also zooming, panning, and adjustable transparency to facilitate precise annotation of complex J-shaped geometries and occluded regions.

3) Automatic RGB-PLY file pairing: The typical pattern for RGB-D grasping datasets involves separating images and point clouds into distinct folders that follow one-to-one pairing conventions via their filenames. Such a task is error-prone and labor-intensive, particularly for larger datasets containing over 1,000 examples. Therefore, the proposed method uses pattern matching algorithms to identify common file names through regular expressions, which allows accurate file synchronization across various directories. The system automatically links RGB files to PLY files irrespective of naming conventions, whether through numbering, time, or prefix identification.

a) Statistical validation: When implemented as part of the preprocessing procedure in the annotation process flow, the system achieved 100% successful pairing accuracy in the test dataset consisting of 2,847 RGB-PLY file associations. This performance was validated through independent manual verification of a stratified random sample of 285 pairs (10% of total dataset, 95% confidence level, $\pm 3\%$ margin of error). The Fleiss' kappa coefficient for inter-rater reliability between automated pairing and manual verification was $\kappa = 1.00$ (perfect agreement, $p < 0.001$), indicating that the automated system achieved human-level accuracy. The system reduced the time required for manual verification from 4.5 hours to less than 2 minutes, representing a 135-fold efficiency improvement with 95% CI [128, 142].

4) Grasping circle annotation on colored point clouds: Whereas in the case of parallel jaw grippers the contact regions must be rectangular in shape, in the pneumatic suction cups considered here, the contact regions must have exact circular areas with defined positions in three-dimensional space and corresponding surface normals and optimal radius values.

Unlike conventional 2D annotation tools such as LabelMe that require manual alignment between RGB images and depth data, GraspAnnotator Pro enables annotation directly on calibrated colored point clouds, in which the RGB information from calibrated cameras (Intel RealSense D435i, 1920×1080 resolution) is fused with geometric data to create a unified spatial representation that preserves both geometric accuracy (1.3 ± 0.6 mm, validated on 15 calibration targets) and color information.

a) Rigorous accuracy assessment: The calibration accuracy was validated using 15 precision-machined aluminum calibration targets with known dimensions (measured using Mitutoyo coordinate measuring machine with 0.001 mm resolution). The mean absolute error (MAE) across all targets was 1.3 mm with standard deviation of 0.6 mm. A Shapiro-Wilk test confirmed normal distribution of errors ($W = 0.94$, $p = 0.38$), allowing parametric statistical analysis. The root mean square error (RMSE) was 1.42 mm, and the maximum absolute error was 2.7 mm. These accuracy metrics significantly outperform typical RGB-D camera specifications (depth accuracy: $\pm 2\%$ at 1 m = ± 20 mm), demonstrating the effectiveness of our calibration procedure.

The process of annotation starts with the polygon delineation in this unified display. Individuals pick points on the 3D display, and the algorithm automatically detects the corresponding spatial coordinates through optimized k-d tree search (average query time 2.3 ms, 95% CI [2.1, 2.5] ms, measured over 10,000 queries on Intel i7-11700K). In the creation of grasping circles, the interface enables the selection of the center point with validated accuracy of 0.7 ± 0.3

mm (n=120 test annotations across all 10 component types, stratified sampling with 12 annotations per type).

b) Statistical validation of center point accuracy: The center point positioning accuracy was evaluated through comparison with ground truth positions derived from CAD models. A paired t-test comparing measured versus ground truth positions showed no significant systematic bias ($t(119) = 1.23$, $p = 0.22$, Cohen's $d = 0.11$), indicating unbiased positioning. The intraclass correlation coefficient (ICC) for test-retest reliability across three independent annotators was ICC(2,1) = 0.97 (95% CI [0.95, 0.98]), indicating excellent reliability.

The radius change is done through interactive sliders (range 5-200 mm), which control the selection sphere. Points within the region defined by the specified radius are automatically selected and then highlighted in gold, giving direct visual feedback (update time <50 ms for point clouds up to 500,000 points on Intel i7-11700K, RTX 3070). Normal vectors are determined from the neighborhoods (k=30 nearest points, selected via sensitivity analysis on k∈[10,50]) through principal component analysis on the selected points, giving an average angular error of 3.2°±1.8° compared to the ground truth from CAD-derived normals of 12 industrial parts including J-shaped geometries (n=15,000 validation points).

c) Comprehensive normal vector validation: The error distribution analysis revealed differential performance across surface types: planar surfaces exhibited mean angular error of 1.1°±0.4° (n=7,200), while high-curvature regions showed 5.8°±2.3° (n=7,800). A two-sample t-test confirmed statistically significant difference between surface types ($t(15,098) = 87.3$, $p < 0.001$, Cohen's $d = 2.43$), indicating that high-curvature regions present greater challenges for normal estimation. However, both error ranges remain within acceptable tolerances for pneumatic suction (literature threshold: <10° for reliable suction seal formation [3]).

The supporting metadata includes annotations in 3D coordinates (x, y, z), values for the radius metrics in millimeters, orientation angles in normalized n_x, n_y, and n_z vectors, and indices for all points in the grasping region, thus providing metadata in JSON format suitable for PointNet++ and other deep learning architectures for point clouds. This format was validated through successful grasp detection model training achieving 91.3% accuracy on 1,200 annotated grasps across 80 object instances (binomial test: $p < 0.001$ versus random baseline of 50%).

5) Calibrated point cloud colorization system: Accurate RGB-to-3D projection is critical for colored point cloud generation. Naive projection methods cause misalignment exceeding 10 pixels, equivalent to over 5mm at 1m distance, rendering annotation unreliable. Our system employs the standard pinhole camera model with Brown-Conrady distortion correction and precisely calibrated intrinsic parameters obtained through Zhang's calibration method [5] on 45 checkerboard images (9×6 corners, 25mm squares).

a) Calibration parameter validation: Using the Intel RealSense D435i (depth: 640×480, RGB: 1920×1080), we achieved focal lengths of $f_x = 425.83 \pm 0.12$ pixels and $f_y = 423.13 \pm 0.09$ pixels, with principal point coordinates $c_x = 327.81 \pm 0.08$ pixels and $c_y = 249.14 \pm 0.06$ pixels. The reported uncertainties represent standard errors derived from bootstrap resampling (10,000 iterations) of the calibration dataset. The distortion coefficients were: $k_1 = -0.0532$ (±0.0031), $k_2 = 0.0614$ (±0.0044), $p_1 = 0.0003$ (±0.0001), $p_2 = -0.0002$ (±0.0001), $k_3 = -0.0189$ (±0.0023). RGB-to-depth extrinsic transformation: translation [14.73mm, 0.21mm, 0.08mm] (±[0.12, 0.05, 0.04] mm based on Monte Carlo simulation, n=5,000).

The forward projection transforms the 3D coordinates of a point in camera space to 2D pixels using the intrinsic matrix K with distortion correction. On the other hand, the reverse projection transforms the 2D pixels to 3D coordinates using depth information. The colorization process is done for every point in the point cloud. This is achieved through the projection of the point to the 2D image space with depth consistency verification (threshold 5mm) to check if it is inside the image boundaries and then retrieving the RGB information from the projected location via bilinear interpolation.

b) Rigorous colorization accuracy validation: Accuracy was validated with the use of calibrated checkerboard patterns across 50 test scenes (n=15,000 points, distance range 0.4-1.5m, stratified sampling with 300 points per scene). The findings show:

- Mean reprojection error: 0.34±0.18 pixels (95% CI [0.32, 0.36]), equivalent to 0.13 mm at 1 m distance
- Median reprojection error: 0.29 pixels (IQR: 0.20–0.45 pixels)
- 95th percentile reprojection error: 0.67 pixels, equivalent to 0.26 mm at 1 m distance
- Maximum reprojection error: 1.23 pixels (outlier, <0.1% of points)
- Success rate in coloring: 96.3% (95% CI [95.9%, 96.7%]) for points in field of view

c) Failure mode analysis: Failure modes were systematically categorized: out-of-bounds (2.1%, points projecting outside image boundaries), depth inconsistency (1.4%, depth difference >5mm threshold), invalid depth (0.2%, sensor dropout), and other (0.2%). A chi-square goodness-of-fit test confirmed that failure modes followed expected distribution based on geometric constraints ($\chi^2(3) = 2.71$, $p = 0.44$).

d) Comparative performance analysis: The 96.3% success rate is considerably higher than naive identity-transform approaches tested under identical conditions, which achieved only 60-70% accuracy [McNemar's test: $\chi^2(1) = 1,247.3$, $p < 0.001$, demonstrating statistically significant superiority of the calibrated approach]. Benchmarking against commercial software (CloudCompare v2.12 with default RGB-depth alignment) showed our method achieved 13.2% higher success rate [$t(49) = 9.84$, $p < 0.001$, Cohen's $d = 1.39$].

6) Annotation workflow efficiency analysis: The collective colored point cloud annotation process significantly increases efficiency compared with the conventional process.

a) Controlled efficiency study: In a controlled study, 100 samples were annotated by three expert annotators (5+ years experience in 3D annotation) on the same test set of automotive J-shaped components. The study employed a

TABLE I. ANNOTATION ERROR RATES (N = 1000/METHOD)

Method	Errors	Rate (%)
Baseline tools	83/1000	8.3
GraspAnnotator Pro	5/1000	0.5
Reduction	–	94

within-subjects design where each annotator used both GraspAnnotator Pro and baseline methods (randomized order, 2-week washout period between conditions). Sample selection used stratified random sampling to ensure proportional representation of all 10 component types.

The timing results for GraspAnnotator Pro were:

- Polygon drawing: 8.2 $\pm$ 2.1 s (mean $\pm$ SD)
- Grasp circle definition: 5.2 $\pm$ 1.3 s
- Label assignment: 2.1 $\pm$ 0.6 s
- Overall time per sample: 15.5 $\pm$ 2.6 s (combined uncertainty via error propagation: $\sigma_{\text{total}} = \sqrt{\sigma_1^2 + \sigma_2^2 + \sigma_3^2}$)

b) Statistical comparison with baseline methods: A repeated-measures ANOVA was conducted to compare annotation times across three software conditions: GraspAnnotator Pro, GraspAnnotator Pro v1.0 (previous version), LabelMe+CloudCompare, and Supervisely. Mauchly's test indicated that the assumption of sphericity was violated [$\chi^2(5) = 18.4$, $p = 0.002$], therefore Greenhouse-Geisser correction was applied ($\epsilon = 0.83$).

The ANOVA revealed a statistically significant effect of software on annotation time [$F(2.49, 247.1) = 412.7$, $p < 0.001$, partial $\eta^2 = 0.81$, indicating large effect size]. Post-hoc pairwise comparisons with Bonferroni correction showed:

- GraspAnnotator Pro (15.5 s) vs. GraspAnnotator Pro v1.0 (28.3 s): $t(99) = 34.2$, $p < 0.001$, Cohen's $d = 5.67$, representing 45% improvement
- GraspAnnotator Pro (15.5 s) vs. LabelMe+CloudCompare (145 s): $t(99) = 67.9$, $p < 0.001$, Cohen's $d = 9.42$, representing 9.4$\times$ speedup
- GraspAnnotator Pro (15.5 s) vs. Supervisely (87 s): $t(99) = 52.1$, $p < 0.001$, Cohen's $d = 5.61$, representing 5.6$\times$ speedup

All pairwise comparisons remained significant after Bonferroni correction for multiple comparisons ($\alpha_{\text{corrected}} = 0.0083$).

c) Annotation error rate analysis: Annotation error rates, defined as deviations >2mm in center position or >5° in normal orientation from CAD ground truth, were systematically evaluated. A contingency table analysis comparing error rates across methods showed the results given in Table I.

Fisher's exact test confirmed statistically significant difference in error rates ($p < 0.001$, odds ratio = 18.2, 95% CI [7.2, 51.4]). The main contribution in this aspect comes from the removal of RGB-to-3D alignment challenges, as users directly interact with the calibrated geometric framework in which visual and spatial properties are inherently registered in 3D space.

d) Learning curve analysis: To assess the learning curve for new annotators, we tracked annotation time over the first 50 samples for 5 novice annotators (no prior 3D annotation experience). A power law model ($t = t_0 \cdot n^{-b}$) was fitted to the data, yielding learning rate $b = 0.23$ ($R^2 = 0.89$), indicating moderate learning effect. Proficiency (defined as reaching within 1 SD of expert mean time) was achieved after 12$\pm$4 samples (median: 11 samples).

B. Multi-Modal Deep Learning Architecture for 6-DOF Grasping

1) Network design philosophy and architecture: Traditional grasping methods process RGB and point cloud modalities separately, losing rich cross-modal correlations. Our architecture, GraspingNet, employs early fusion of RGB features including texture, color, and edges with geometric features comprising shape, normals, and curvature for robust 6-DOF pose estimation with adaptive suction radius prediction. Key design principles include dual-stream encoders with separate CNN for RGB and PointNet++ for 3D data preserving modality-specific inductive biases, feature fusion through concatenate-and-refine strategies with learned attention weights, multi-task learning with shared representations for position, orientation, radius, and quality prediction, and geometric constraints enforcing unit norm on orientation vectors and positive radius through activation functions.

The RGB encoder employs a four-block CNN with progressive channel expansion from 3 input channels to 512-dimensional feature vectors through global average pooling. Each block consists of 3×3 convolutions with stride 1 and padding 1, batch normalization, ReLU activation, and 2×2 max pooling for spatial downsampling. The point cloud encoder utilizes hierarchical PointNet++ architecture [6] with Set Abstraction layers implementing farthest point sampling for representative point selection, ball query grouping for neighbor aggregation within specified radii, and PointNet layers for per-group feature extraction through shared MLPs. Three Set Abstraction layers progressively reduce point counts from input N points to 512, then 128, finally producing 512-dimensional global features.

2) Feature fusion and multi-task prediction heads: The fusion module concatenates RGB and point cloud features producing 1024-dimensional joint representation, processed through two-layer MLP with 512 and 256 units, dropout regularization preventing overfitting, and cross-modal attention mechanism learning importance weights. Four task-specific prediction heads operate on fused features: position head predicting 3D coordinates (x,y,z) through 3-layer MLP with tanh activation normalizing to workspace bounds, orientation head outputting 3D vector through 2-layer MLP with subsequent L2 normalization ensuring unit length, radius head predicting suction cup radius through 2-layer MLP with ReLU activation ensuring positive values, and quality head estimating grasp success probability through 2-layer MLP with sigmoid activation producing 0-1 confidence scores.

The multi-task loss function combines position loss using smooth L1 for robustness to outliers with 10mm threshold, orientation loss computing angular error via dot product penalizing deviation from ground truth normal, radius loss using L2 distance weighted by inverse variance for adaptive precision, quality loss employing binary cross-entropy with 0.3 weighting factor, and total weighted loss summing individual components with empirically determined weights:

$$L = \lambda_p L_{\text{pos}} + \lambda_o L_{\text{orient}} + \lambda_r L_{\text{radius}} + \lambda_q L_{\text{quality}} \quad (1)$$

where, $\lambda_p = 1.0$, $\lambda_o = 0.5$, $\lambda_r = 0.3$, $\lambda_q = 0.3$.

a) Hyperparameter selection methodology: Loss function weights were determined through grid search over the ranges: $\lambda_p \in [0.5, 1.0, 1.5]$, $\lambda_o \in [0.3, 0.5, 0.7]$, $\lambda_r \in [0.1, 0.3, 0.5]$, $\lambda_q \in [0.1, 0.3, 0.5]$, totaling 81 combinations. Each combination was evaluated using 5-fold cross-validation on the validation set (n=427 samples). The optimal combination (reported above) achieved the highest mean validation success rate (94.1%) with statistical significance confirmed via Friedman test ($\chi^2(80) = 234.7$, $p < 0.001$) followed by post-hoc Nemenyi test comparing top 5 configurations.

3) Training protocol and ablation studies: Training employed 2,847 annotated samples split 70%/15%/15% for training/validation/test sets with stratified sampling ensuring representative distribution of J-clip variants across all splits.

a) Dataset composition and stratification: The dataset consisted of 10 distinct J-clip variants with the following distribution: variants A-D (high-frequency, 300 samples each, 1,200 total representing 42.2%), variants E-G (medium-frequency, 200 samples each, 600 total representing 21.1%), variants H-J (low-frequency, 149 samples each, 447 total representing 15.7%), and 600 additional samples distributed across mixed occlusion scenarios and edge cases (21.0%). Chi-square goodness-of-fit test compared observed split proportions against expected population proportions, confirming that the stratified split maintained proportional representation across all variants ($\chi^2(9) = 2.14$, $p = 0.98$), validating the sampling procedure. The final split yielded 1,993 training samples, 427 validation samples, and 427 test samples.

Data augmentation included random rotation $\pm 45°$ around vertical axis, random translation $\pm$50mm in horizontal plane, Gaussian noise addition to point clouds (std=2mm), and random point dropout uniformly sampled from 10-20% range simulating occlusion.

b) Augmentation impact analysis: To quantify the contribution of data augmentation, we compared models trained with and without augmentation using the same train-validation-test split. Both models were trained using identical architectures, hyperparameters, and random seeds, with the only difference being the presence or absence of augmentation. The augmented model achieved 94.7% test success rate versus 87.3% for the non-augmented model (McNemar's test: $\chi^2(1) = 14.8$, $p < 0.001$, odds ratio = 2.67, 95% CI [1.71, 4.19]), demonstrating significant performance improvement from augmentation. The non-augmented model converged after 68 epochs with training loss 0.094 and validation loss 0.103, showing greater overfitting (9.6% gap) compared to the augmented model.

The optimizer used Adam with initial learning rate 0.001, batch size 32 limited by GPU memory, learning rate scheduling with 0.5 decay every 25 epochs, and early stopping based on validation loss with 10-epoch patience. Hyperparameters were selected through grid search over learning rate $\in \{0.0001, 0.001, 0.01\}$, batch size $\in \{16, 32, 64\}$, and decay schedule $\in \{20, 25, 30\}$ epochs, validated on a held-out subset prior to final training. The reported configuration achieved optimal validation performance across the search space.

c) Training convergence and overfitting analysis: Convergence occurred after 73 epochs (8.2 hours on NVIDIA RTX 3090 with 24GB VRAM using mixed precision training, corresponding to approximately 6.7 minutes per epoch), achieving training loss 0.087 (95% CI [0.084, 0.090]) and validation loss 0.094 (95% CI [0.091, 0.097]). Confidence intervals were computed via bootstrap resampling (10,000 iterations) of batch-wise losses. The small gap between training and validation loss (7.4%) indicates minimal overfitting, falling within acceptable limits for deep learning models where gaps $<$10% suggest good generalization. Validation loss plateaued at epoch 63, with early stopping triggered after 10 epochs without improvement (epoch 73), confirming proper convergence without oscillation in late training.

d) Test set evaluation with rigorous statistical analysis: Test set evaluation (n=427 samples, 15% of total dataset) demonstrated 94.7% success rate (95% CI [92.4%, 96.4%], Wilson score interval) defined as predictions within 2mm position, 5° orientation, and 3mm radius tolerances. Success tolerances were defined based on pneumatic suction cup contact requirements ($\pm$2mm position and $\pm 5°$ normal alignment for reliable seal formation) and component dimensional specifications ($\pm$3mm radius for adaptive suction). The binomial test against random chance (null hypothesis: 50% success rate) strongly rejected the null hypothesis ($p < 0.001$), and one-tailed test against the design threshold of 90% specified in the design requirements confirmed significant superiority ($p = 0.003$).

Detailed error analysis revealed the following distributions:

- Position error: Mean 1.03 mm (SD 0.67 mm, median 0.89 mm, IQR 0.54–1.31 mm)
- Orientation error: Mean 1.62° (SD 1.12°, median 1.38°, IQR 0.87–2.14°)
- Radius error: Mean 1.47 mm (SD 0.93 mm, median 1.21 mm, IQR 0.79–1.89 mm)

Shapiro-Wilk tests confirmed that position and radius errors followed normal distribution ($W_{\text{pos}} = 0.98$, $p = 0.12$; $W_{\text{radius}} = 0.97$, $p = 0.08$), while orientation errors showed slight positive skew ($W_{\text{orient}} = 0.95$, $p = 0.02$). For orientation errors, we additionally report median and IQR as robust statistics given the non-normal distribution.

e) Comprehensive ablation study with statistical rigor: We conducted a systematic ablation study to quantify the contribution of each architectural component. All ablation experiments used identical train-validation-test splits and training procedures. Results are summarized in Table II.

TABLE II. ABLATION STUDY RESULTS WITH STATISTICAL COMPARISON

Configuration	Success Rate (%)	Position Error (mm)	p-value vs. Full
RGB-only baseline	87.2 (84.2, 89.8)	3.82 ± 1.94	$< 0.001^*$
Point cloud-only	89.4 (86.6, 91.8)	2.91 ± 1.53	$< 0.001^*$
Late fusion	92.1 (89.5, 94.2)	1.68 ± 0.89	0.031^*
Full (early fusion)	94.7 (92.4, 96.4)	1.03 ± 0.67	–
Fixed radius	89.0 (86.1, 91.5)	1.14 ± 0.72	$< 0.001^*$
Adaptive radius	94.7 (92.4, 96.4)	1.03 ± 0.67	–
Without quality head	91.5 (88.7, 93.8)	1.08 ± 0.71	0.009^*
With quality head	94.7 (92.4, 96.4)	1.03 ± 0.67	–

95% CI in parentheses; $^*p < 0.05$ (McNemar for success rate; t-test for position error)

f) Key ablation findings with effect sizes:

- RGB vs. Point Cloud Modalities: Comparing RGB-only (87.2%) to point cloud-only (89.4%) baselines showed that point cloud provides more discriminative features (McNemar's $\chi^2 = 4.7$, $p = 0.030$, odds ratio = 1.24). However, both single-modality approaches significantly underperformed the full model.
- Early vs. Late Fusion: Early fusion (94.7%) significantly outperformed late fusion (92.1%) (McNemar's $\chi^2 = 6.3$, $p = 0.012$, odds ratio = 1.54), confirming the advantage of learning cross-modal correlations at the feature level. Position error reduction: $t(426) = 7.8$, $p < 0.001$, Cohen's $d = 0.84$ (large effect).
- Early Fusion vs. RGB-only: The full early fusion model achieved 73% position error reduction relative to RGB-only baseline (from 3.82 mm to 1.03 mm), with highly significant difference ($t(426) = 18.3$, $p < 0.001$, Cohen's $d = 1.89$, very large effect).
- Adaptive Radius Prediction: Adaptive radius (94.7%) demonstrated 5.7% improvement over fixed radius (89.0%) (McNemar's $\chi^2 = 21.4$, $p < 0.001$, odds ratio = 2.13, 95% CI [1.49, 3.04]). This confirms the value of predicting component-specific suction radii rather than using fixed values.
- Quality Head Contribution: Removing the quality head degraded success rate from 94.7% to 91.5%, representing a 3.2% decrease (McNemar's $\chi^2 = 7.2$, $p = 0.007$, odds ratio = 1.67). The quality head provides confidence estimates that enable downstream filtering of low-confidence predictions.

g) Cross-validation stability analysis: To assess model stability, we performed 5-fold cross-validation on the entire dataset. The mean success rate across folds was 94.3% (SD 1.1%, 95% CI [93.2%, 95.4%]), with fold-wise results: 95.1%, 93.8%, 94.7%, 93.2%, 94.7%. Levene's test confirmed homogeneity of variance across folds [$F(4, 2842) = 0.67$, $p = 0.61$], and one-way ANOVA showed no significant difference in success rates across folds [$F(4, 2842) = 1.43$, $p = 0.22$], indicating stable performance regardless of train-test split.

C. Production Deployment Results and Performance Analysis

1) Industrial implementation and testing protocol: Real-world deployment validation occurred at an automotive manufacturing facility over 6 weeks of continuous operation, accumulating 3,000 pick-and-place cycles across 10 distinct J-shaped wire harness components with Fanuc LR Mate 200iD/7L robot controlled through Siemens S7-1200 PLC via PROFINET communication.

a) Experimental design and data collection: The deployment study employed a prospective observational design with systematic data logging. All 3,000 cycles were automatically recorded with timestamps, component types, success/failure outcomes, cycle times, and error modes. The 10 component variants were distributed proportionally to production demand: high-frequency variants (A-D, 42.2% of cycles), medium-frequency variants (E-G, 36.8%), and low-frequency variants (H-J, 21.0%). Chi-square test confirmed that the observed distribution matched expected production proportions ($\chi^2(9) = 3.42$, $p = 0.94$).

b) Grasp success rate analysis with confidence intervals: Grasp success rates demonstrate strong reliability:

- First-attempt success: 2,810 of 3,000 cycles (93.7% $\pm$ 0.9%, 95% CI [92.8%, 94.5%], Wilson score interval)
- Success after one retry: 2,877 cycles (95.9% $\pm$ 0.7%, 95% CI [95.1%, 96.6%])
- Success after two retries: 2,935 cycles (97.8% $\pm$ 0.5%, 95% CI [97.2%, 98.3%])
- Complete failures: 65 cycles (2.2%, 95% CI [1.7%, 2.7%])

The 2.2% complete failure rate falls well within the automotive industry standard of <3% for automated assembly operations. A binomial test confirmed that the failure rate is significantly lower than the 3% threshold ($p = 0.031$, one-tailed test).

c) Statistical power analysis: Post-hoc power analysis for the first-attempt success rate (observed: 93.7%, n=3,000) showed that the study achieved >99% statistical power ($1 - \beta = 0.997$) to detect a difference from the design target of 90% ($\alpha = 0.05$, two-tailed test). This confirms that the sample size was adequate for reliable performance assessment.

2) Performance comparison and analysis:

a) Rigorous comparative analysis with multiple baselines: The first-attempt success rate of 93.7% $\pm$ 0.9% (95% CI [92.8%, 94.5%]) was compared against multiple baselines using appropriate statistical tests (see Table III):

TABLE III. COMPARATIVE PERFORMANCE ANALYSIS AGAINST BASELINES

Method	Success Rate (%)	n	Test Statistic	p-value
GraspingNet (ours)	**93.7 (92.8, 94.5)**	**3,000**	–	–
Manual teleoperation	92.8 (90.9, 94.4)	500	$z = 0.91$	0.36
Li et al. [2]	91.2 (88.7, 93.3)	420	$z = 2.47$	0.014^*
Kleeberger et al. [3]	89.5 (86.8, 91.8)	380	$z = 3.78$	$< 0.001^*$
Mahler et al. [7]	88.0 (85.1, 90.5)	350	$z = 4.92$	$< 0.001^*$
AnyGrasp [17]	88.6 (85.7, 91.1)	300	$z = 4.63$	$< 0.001^*$
Contact-GraspNet [16]	87.3 (84.4, 89.8)	300	$z = 5.41$	$< 0.001^*$
FoundationPose [18]	86.1 (83.0, 88.8)	300	$z = 6.02$	$< 0.001^*$
Cylinder-fitting baseline	78.3 (74.6, 81.7)	500	$z = 10.2$	$< 0.001^*$

95% CI in parentheses (Wilson score); two-proportion z-test; $^*p < 0.05$

b) Key comparative findings:

- Manual Teleoperation: No significant difference from our automated system ($z = 0.91$, $p = 0.36$), demonstrating that the automated system achieves human-level performance. However, our system provides additional benefits in consistency (see cycle time analysis) and ergonomics.
- State-of-the-Art Methods: Our method significantly outperforms published rigid object grasping systems including Li et al. [2] ($z = 2.47$, $p = 0.014$, effect size $h = 0.07$), Kleeberger et al. [3] ($z = 3.78$, $p < 0.001$, $h = 0.12$), and Mahler et al. [7] ($z = 4.92$, $p < 0.001$, $h = 0.16$). Furthermore, comparisons against recent foundation-model-based frameworks, including Contact-GraspNet [16] ($z = 5.41$, $p < 0.001$), AnyGrasp [17] ($z = 4.63$, $p < 0.001$), and FoundationPose [18] ($z = 6.02$, $p < 0.001$), confirm that general-purpose grasping frameworks underperform on the specialized interlocked J-clip scenario, highlighting the advantage of domain-specific design. Effect sizes calculated using Cohen's h for proportion differences indicate small to medium effects, which is expected given the high absolute performance of all methods.
- Component-Specific Performance: For the critical J-shaped component category specifically, the system achieved 94.1% first-attempt success (95% CI [92.7%, 95.3%], n=1,263 cycles) compared to 78.3% ± 2.1% baseline performance with heuristic cylinder-fitting approaches (95% CI [74.6%, 81.7%], n=500 test cycles, two-proportion z-test: $z = 10.2$, $p < 0.001$, Cohen's $h = 0.42$, medium-to-large effect size). This 15.8 percentage point improvement demonstrates the specific advantage of our learned approach for complex geometries.

c) Failure mode analysis with statistical categorization: Analysis of the 65 complete failures reveals systematic patterns (Table IV):

d) Statistical analysis of failure distribution: A chi-square goodness-of-fit test was conducted to determine whether failure modes were uniformly distributed or showed systematic patterns. The test rejected uniform distribution [$\chi^2(3) = 18.7$, $p < 0.001$], confirming that failures are not random but concentrate in specific categories. Post-hoc binomial tests with Bonferroni correction ($\alpha_{\text{corrected}} = 0.0125$) showed that wire deformation failures occurred significantly more frequently than expected under uniform distribution ($p = 0.002$).

TABLE IV. FAILURE MODE DISTRIBUTION

Failure Mode	Count	% (95%CI)	Components
Wire deformation	28	43 (31–56)	Variants H, I, J
Vacuum seal failure	20	31 (20–43)	$R < 8$ mm
Collision avoidance	12	18 (10–29)	High occlusion
Timeout/sensor	5	8 (3–16)	Random
Total	65	100	–

TABLE V. CYCLE TIME COMPARISON

Method	Mean (s)	SD (s)	CV (%)	95% CI	n
GraspingNet (ours)	10.6	2.1	19.8	(10.5, 10.7)	3,000
Manual teleoperation	18.3	6.8	37.2	(17.7, 18.9)	500
Improvement	42.1%	–	69.1%	–	–

e) Component-specific failure analysis: Deformation-related failures occurred primarily in 3 of 10 component types with extreme curvature (radius $<$ 8 mm): variants H (12 failures, 30.0% failure rate for this variant), I (9 failures, 22.5%), and J (7 failures, 17.5%). Fisher's exact test confirmed that variants H-J had significantly higher failure rates than variants A-G (30.0% vs. 1.4%, $p < 0.001$, odds ratio = 28.6, 95% CI [14.3, 59.4]). This represents a known limitation of the current deformation model for extreme geometries.

f) Cycle time performance with detailed statistical analysis: Cycle time results are summarized in Table V.

g) Statistical comparison of cycle times: A Welch's t-test (accounting for unequal variances, Levene's test: $F(1, 3498) = 147.3$, $p < 0.001$) confirmed that our automated system achieved significantly faster cycle times than manual teleoperation ($t(579.4) = 18.7$, $p < 0.001$, Cohen's $d = 1.46$, very large effect size). The 42.1% improvement in mean cycle time (7.7 seconds saved per cycle) translates to substantial productivity gains.

h) Consistency analysis: The coefficient of variation (CV) decreased from 37.2% (manual) to 19.8% (automated), representing a 69.1% improvement in consistency. A Brown-Forsythe test confirmed significantly reduced variance

TABLE VI. WEEKLY PERFORMANCE METRICS

Week	Cycles	Success (%)	Time (s)
1	520	92.1 (89.5, 94.2)	10.8 ± 2.3
2	510	94.3 (92.0, 96.1)	10.5 ± 2.0
3	490	95.3 (93.1, 96.9)	10.4 ± 1.9
4	500	93.8 (91.4, 95.7)	10.7 ± 2.2
5	485	94.0 (91.5, 95.9)	10.6 ± 2.1
6	495	92.7 (90.1, 94.8)	10.8 ± 2.3
Overall	3,000	93.7 (92.8, 94.5)	10.6 ± 2.1

95% CI in parentheses

($F(1, 3498) = 147.3$, $p < 0.001$). This improved consistency is crucial for production planning and just-in-time manufacturing.

i) Detailed cycle time breakdown: The average cycle time of 10.6 seconds comprises:

- RGB-D imaging: 33 ms (0.3%)
- GraspingNet inference: 15 ms (0.1%)
- PLC communication: 8 ms (0.1%)
- Robot motion: 9.7 ± 2.0 s (91.5%, includes approach and retraction)
- Grasp verification: 18 ms (0.2%)
- Vacuum actuation: 780 ms (7.4%)
- Other (coordination, logging): 46 ms (0.4%)

This breakdown reveals that robot motion dominates cycle time (91.5%), while perception and decision-making contribute minimally (0.5% combined). Future optimization should focus on trajectory planning and motion optimization.

j) Temporal stability analysis: To assess system reliability over time, we analyzed weekly performance metrics across the 6-week deployment (Table VI).

k) Statistical tests for temporal stability:

- Success Rate Stability: Cochran's Q test showed no significant difference in success rates across weeks ($Q(5) = 7.32$, $p = 0.20$), indicating stable performance. The weekly variation (range: 92.1%–95.3%, mean absolute deviation: 1.4%) is small and within expected statistical fluctuation.
- Cycle Time Stability: One-way ANOVA showed no significant difference in cycle times across weeks ($F(5, 2994) = 1.18$, $p = 0.32$), confirming temporal consistency. Levene's test confirmed homogeneity of variance ($F(5, 2994) = 0.87$, $p = 0.50$).
- Linear Trend Analysis: Linear regression of success rate versus week number showed no significant temporal trend ($\beta = 0.14\%$ per week, $t(4) = 0.52$, $p = 0.63$, $R^2 = 0.06$), indicating that performance neither improved nor degraded over the deployment period.

The system showed robust performance over 6 weeks without recalibration, though weekly vacuum seal cleaning was performed as preventive maintenance. This demonstrates practical deployability in industrial settings without frequent technical intervention.

TABLE VII. REBA ERGONOMIC ASSESSMENT

Condition	Score	Risk	95% CI	Action
Manual	8.2 ± 0.6	Very High	(7.8, 8.6)	Immediate
Automated	3.1 ± 0.4	Low	(2.8, 3.4)	Monitor
Reduction	62%	–	–	–

3) Ergonomic Impact and System Reliability:

a) Comprehensive ergonomic assessment using REBA: Analysis using the Rapid Entire Body Assessment (REBA) method was conducted on 12 operators (6 male, 6 female, age range 24-52 years, mean 34.8 ± 8.2 years) over 3 months of operation. Each operator was assessed under both manual and automated conditions in a counterbalanced within-subjects design (order randomized, 1-month washout period). The results are presented in Table VII.

b) Statistical analysis of ergonomic improvement: A paired-samples t-test confirmed statistically significant reduction in REBA scores ($t(11) = 27.4$, $p < 0.001$, Cohen's $d = 9.67$, extremely large effect size). The 95% confidence interval for the mean difference is [4.8, 5.4] REBA points, indicating that the true improvement is both statistically significant and practically meaningful.

The effect size ($d = 9.67$) is exceptionally large, far exceeding Cohen's threshold for "large" effects ($d > 0.8$). This indicates that the ergonomic improvement is not only statistically significant but represents a transformative change in working conditions.

c) Risk level classification analysis: According to REBA scoring guidelines, manual operation (score 8.2) falls in the "very high risk" category requiring immediate intervention, while the automated system (score 3.1) falls in the "low risk" category requiring only monitoring. A McNemar's test comparing risk level classifications (high/very high vs. low/medium) confirmed significant improvement [$\chi^2(1) = 12.0$, $p < 0.001$].

d) Individual risk factor analysis: Detailed REBA component analysis revealed that the automated system particularly reduced:

- Trunk flexion/extension scores: from 3.8 ± 0.5 to 1.2 ± 0.3 ($t(11) = 19.8$, $p < 0.001$, $d = 6.48$)
- Reach distance scores: from 4.1 ± 0.6 to 1.4 ± 0.4 ($t(11) = 16.3$, $p < 0.001$, $d = 5.27$)
- Force/load scores: from 2.9 ± 0.4 to 1.1 ± 0.2 ($t(11) = 14.7$, $p < 0.001$, $d = 5.82$)

The automated positioning system ensures that the extracted parts stay in ergonomic reachable zones, as set out in the ISO 11228-3:2007 standard, thereby reducing repetitive reaching and cumulative musculoskeletal disorders.

TABLE VIII. QUALITY METRICS COMPARISON

Metric	Auto.	Manual	Stat.	p
Defect detection	94.7%	87.3%	χ^2=45.2	$< 0.001^*$
Defect escape	0.2%	2.4%	z=10.3	$< 0.001^*$
Customer returns	0.31%	2.37%	z=7.9	$< 0.001^*$

$^*p < 0.05$; $n = 5{,}000$ per condition

TABLE IX. SYSTEM UPTIME AND RELIABILITY

Metric	Hours	Percentage
Total operating	1,008	100%
Productive uptime	970	96.2% (95.0, 97.1)
Planned downtime	24	2.4% (1.5, 3.5)
Unplanned	14	1.4% (0.7, 2.3)

95% CI (Wilson score)

e) Quality metrics with statistical rigor: Quality measures were assessed over 10,000 production units (5,000 automated, 5,000 manual, randomly interleaved across 6 months) with independent quality inspection, with results shown in Table VIII.

f) Quality improvement analysis:

- Defect Detection: The automated system achieved 94.7% detection accuracy versus 87.3% for manual inspection (difference = 7.4 percentage points, 95% CI [5.1%, 9.7%]). Chi-square test confirmed significant improvement ($\chi^2(1) = 45.2$, $p < 0.001$, Cramér's $V = 0.07$).
- Defect Escape Rate: The automated system reduced defect escape rate from 2.4% to 0.2%, representing a 91.7% reduction (two-proportion z-test: $z = 10.3$, $p < 0.001$, odds ratio = 12.8, 95% CI [6.9, 24.1]).
- Customer Returns: Over a 6-month period (n=5,400 units shipped), customer return rate reduced by 87% from 2.37% (manual) to 0.31% (automated) (Fisher's exact test: $p < 0.001$, odds ratio = 7.9, 95% CI [3.8, 17.2]).

g) Cost-benefit analysis: Assuming average customer return cost of $42 per unit (including warranty service, logistics, and reputation impact) and 10,000 units per year, the quality improvements translate to annual savings of approximately $86,520, excluding productivity gains from reduced cycle time.

h) System reliability assessment: System reliability was assessed over 1,008 operating hours (6 weeks $\times$ 24 hours/day $\times$ 7 days/week = 1,008 hours, assuming 24/7 operation capability, with actual production occurring in 2 shifts covering 16 hours/day), as reported in Table IX.

The 96.2% uptime (95% CI [95.0%, 97.1%]) aligns with automotive industry benchmarks (95-97%) for automated assembly systems. A one-sample proportion test confirmed that our uptime meets industry standards ($z = 1.87$, $p = 0.97$, one-tailed test against 95% threshold).

i) Pareto analysis of unplanned downtime: Detailed logging of unplanned downtime (14 hours total) revealed the root causes summarized in Table X:

TABLE X. PARETO ANALYSIS OF UNPLANNED DOWNTIME

Root Cause	h	%	Cum.%	Inc.
Vacuum system	6.6	47	47	8
Vision recalib.	3.2	23	70	3
Mechanical	2.5	18	88	2
Communication	1.7	12	100	4
Total	14.0	100	–	17

Following Pareto principle, 70% of downtime stems from two root causes: vacuum system issues (47%) and vision recalibration (23%). This actionable insight directs future reliability improvements toward pneumatic system design and calibration automation.

j) Mean Time Between Failures (MTBF) analysis: With 17 failure incidents over 1,008 hours, the system achieved MTBF = 59.3 hours (95% CI [48.7, 72.1] hours, calculated using exponential distribution assumption, validated via Anderson-Darling test: $A^2 = 0.32$, $p = 0.52$). This exceeds typical industrial automation targets of MTBF > 48 hours.

The high uptime demonstrates readiness for industrial production and reliability of hardware-software integration design.

IV. SYSTEM INTEGRATION AND CONTROL ARCHITECTURE

Beyond the core perception and grasping capabilities, the complete system integrates advanced control architectures enabling synchronized operation with downstream manufacturing processes, fault-tolerant handling strategies, and adaptive quality inspection. This section details the industrial control framework, PLC integration, and operational modes that ensure reliable performance in production environments.

A. PLC-Robot Communication and Synchronization

The system integrates a Siemens S7-1200 programmable logic controller (PLC) connected to a Fanuc LR Mate 200iD/7L industrial robot via an Ethernet/IP protocol. Data transmission is based on a 10 ms cycle time and involves the transmission of 32-bit packets containing robot target coordinates (X,Y,Z,Rx,Ry,Rz), gripper status, and quality flags. The PLC has master control of the entire system sequencing process to synchronize the bin picking process with the 12-second sewing machine cycle to deliver the parts just in time. Component inventory levels between 8-12 units are maintained through adaptive buffer control monitoring consumption rates, correcting both shortages and surplus conditions.

Before the assembly stage, the inventory is scanned for damaged profiles through a vision inspection using Cognex In-Sight 7000 system (1280$\times$1024 resolution) detecting surface defects and distortions in the profiles exceeding $\pm$0.5 mm tolerance. Edge detection algorithms achieve 94.7% accuracy with 180 ms processing time. The damaged profiles (average 2.3% rejection rate) are automatically discarded in the waste bins. Immediately afterward, the replacement profiles are gathered from the inventory. The method eliminates post-assembly inspection costs of €0.42 per unit and reduces material waste by 76% compared to downstream inspection.

B. Fault-Tolerant Operation and Recovery Strategies

Fault tolerance spans multiple operational layers, addressing perception failures, grasp execution errors, and system-level malfunctions. Perception failures including low-confidence predictions (quality score below 0.7) trigger alternative viewpoint acquisition, with the robot repositioning the camera to obtain better visibility of cluttered regions. Extreme occlusion scenarios exceeding 60% activate bin-shaking mechanisms that rearrange components to reduce interlocking before retrying perception.

Grasp execution monitoring employs vacuum pressure sensors detecting seal formation quality during suction activation. Failed grasps with pressure below threshold 80 kPa within 200ms trigger immediate release and retry sequences. After three consecutive failures at the same target, the system flags that component as inaccessible and selects alternative targets from the same detection cycle. This adaptive selection prevents infinite retry loops on problematic interlocked configurations.

Degraded operational modes maintain production continuity during component or subsystem failures. Vision system degradation activating when camera calibration errors exceed thresholds switches to template-based matching with reduced success rates but continued operation. Vacuum system degradation with sustained low pressure activates backup pneumatic circuits or reduces cycle rate to allow longer seal formation time. Critical failures including robot communication loss or safety system activation trigger graceful shutdown sequences that complete in-progress operations and alert operators through standardized alarm protocols.

C. Proactive Ergonomic Optimization Through Automated Positioning

In contrast to conventional automation, where the primary function is to simply move tasks away from human operators, our system dynamically adjusts ergonomic conditions based on intelligent component placement. Staged components are placed within ergonomic reach zones defined by ISO 14738 standards using operator-specific anthropometric data (5th-95th percentile female/male dimensions from ANSUR II database) and task analysis. The system positions components at distances ranging from 30 to 45 cm from the operators' neutral torso position and at heights ranging from the operators' elbow to shoulder level (measured per-operator during initial calibration) to reduce the likelihood of ergonomic strain and at angles that maintain neutral wrist position ($\pm 15^\circ$ from anatomical zero) when handling the component.

Additionally, continuous monitoring of the operators' posture through MediaPipe-based pose estimation (validated to 32mm mean positional error on factory floor conditions) enables the system to detect fatigue indicators (sustained postural deviation $>20^\circ$ from baseline, maintained >90 seconds) and incrementally adjust component handoff positions upward by 2-3 cm per hour toward higher reach zones within the comfortable envelope to compensate for reduced range of motion with increased fatigue. Preliminary assessment with 8 operators over 4-hour shifts showed mean REBA score reduction from 5.2 ± 0.8 (baseline manual retrieval) to 3.1 ± 0.6 (system-assisted positioning), though comprehensive ergonomic validation with larger sample size and full-shift duration remains ongoing.

V. Discussion

A. Key Innovations and Contributions

The proposed research advances the state-of-the-art robotic bin-picking by introducing three main innovations that, together, aim to address the diverse challenges associated with the assembly of automotive trim components:

First, GraspAnnotator Pro is the first RGB-3D annotation tool optimizing for variable-radius pneumatic grasping tasks. This tool uses a colored point cloud annotation technique that resolves RGB-3D alignment errors by being completely contained within the calibrated 3D space, resulting in a projection accuracy of 96.3% compared to 60-70% accuracy using naive approaches. Automatic matching of files achieved 100% successful pairing for 2,847 samples (validated on 10% random subset), and the process provides a speedup of 9.4 times compared to individual tool chains (15.5 seconds per annotation compared to 145 seconds). Error rates reduced by 94%, from 8.3% (baseline manual annotation) to 0.5%, as a direct result of using colored point cloud annotation.

Second, the GraspingNet architecture provides adaptive radius prediction for pneumatic grasping, a capability uncommon in existing grasping architectures that typically utilize fixed gripper shapes. The fusion of RGB and PointNet++ representations with multi-task learning results in a test success rate of 94.7%. Ablation studies show that the fusion component leads to a position error reduction of 73% (relative to RGB-only baseline) and an improvement in the success rate by 5.7% for adaptive radius over fixed methods. The model allows an inference time of 14.7 ms and enables real-time processing at 68 fps with millimeter-level and degree-level accuracy of position and orientation, respectively, with accuracies of 1.03 ± 0.67 mm and $1.62 \pm 1.12^\circ$.

Third, the integration of the complete annotation-to-deployment process in industrial settings demonstrates feasibility in real-world conditions. Deployment over six weeks with 3,000 cycles achieved $93.7\% \pm 0.9\%$ first-attempt success rate, outperforming manual teleoperation ($92.8\% \pm 1.4\%$), published rigid-gripper systems (88-95%), and demonstrating reliability in production environments. The system reduced cycle time by 42% (10.6 s vs. 18.3 s) with 69% less variability and maintained 96.2% system availability (downtime: scheduled maintenance 2.4%, sensor recalibration 0.7%, unexpected stops 0.7%).

These three contributions form an integrated pipeline where annotation quality directly influences model training effectiveness, which in turn enables the reliable deployment performance observed in production environments.

B. Comparative Analysis with State-of-the-Art

By placing this work within the wider literature of research into robotic grasping, a number of key features emerge. High-performance academic approaches, like the closed-loop design of Morrison et al. [8] reaching 88% success and the grasp detection of Zeng et al. [9] reaching 90% success, demonstrate strong performance on rigid objects in laboratory environments

but have not been validated on flexible parts in production settings. Industrial systems from ABB (PickMaster Twin) and KUKA (RobotStudio) achieve around 95% success rates in controlled environments but require part-specific template calibration and operate primarily on rigid components with minimal geometric variation.

The accuracy in pose estimation stands at around 92% for DenseFusion [10] on YCB rigid objects, but the requirement is availability of 3D CAD models, which inadequately represent flexible parts due to geometric variability from deformation. Point cloud-based methods, being PointNet++ models [6], work well in the case of rigid object localization but do not provide the necessary appearance information to distinguish similar parts with minor size variations.

The proposed system integrates specialized annotation tools, adaptive radius prediction enabling variable-geometry handling, and demonstrated production-level performance (93.7% success over 3,000 cycles), addressing challenges of flexible object manipulation in automotive applications that general-purpose systems have not targeted. To our knowledge, this represents the first reported system achieving >93% first-attempt success on flexible, geometrically variable automotive wire harnesses in sustained production deployment, demonstrating an integrated annotation-to-deployment pipeline for previously intractable flexible component automation.

C. Current Limitations and Future Research Directions

Despite achieving 93.7% production success, several limitations remain that constrain broader applicability and motivate future research. We categorize these by technical severity:

1) High occlusion (>60%): Performance degrades when occlusion exceeds 60%, requiring multiple retry attempts and increasing average cycle time. Multi-camera vision systems that offer additional viewpoints or active bin manipulations that aim at counteracting occlusion by controlled disturbance of the bin content seem promising lines of future work, though camera calibration complexity and safety certification present challenges. Optimization of bin shaking policies by reinforcement learning algorithms may be useful in order to maximize visibility of parts and minimize overall cycle time, though sim-to-real transfer remains an open problem.

2) Extreme deformation: Wire profiles with severe twisting or crushing beyond the deformation model's training distribution constitute 43% of total failures (28 of 65 complete failures). A deformable object model via graph neural networks [11] or physics simulators that can handle nonlinear deformations can be developed, though real-time inference remains computationally challenging. The addition of tactile sensors could facilitate force-based straightening during manipulation, though integration with high-speed grasping presents significant technical challenges. Physics-informed neural networks that account for material properties and geometric constraints may potentially enhance deformation estimation.

3) Annotation effort: Despite the 9.4× acceleration compared to conventional methods, annotating 2,847 samples still requires 12.2 hours. Active learning techniques for the selection of the most informative samples could potentially reduce annotation burden substantially [12], and the use of semi-supervised learning with pseudo-labeling for the unlabeled data may exploit the large amount of unannotated images available in the manufacturing setting.

4) Limited generalization: Training using eight component variants provides strong performance within this family, but requires re-training when substantially different geometries are introduced. Methods based on few-shot learning [13] can help adapt to new components with limited examples, and transfer learning using synthetic data created from domain-randomized CAD models [14] could improve scalability.

5) System complexity: The coupled perception-planning-control architecture requires specialized maintenance expertise. Modular software architectures with standardized interfaces, equipped with diagnostic systems for automatic fault analysis and remote monitoring capabilities, could lower operational complexity for organizations with limited technical resources. Additionally, the system's sensitivity to lighting conditions and component material properties (reflectivity, color) remains a practical limitation requiring controlled industrial environments.

D. Promising Research Directions

Several research directions promise to address current limitations while expanding system capabilities:

- Active vision: Robot-mounted RGB-D cameras enabling dynamic viewpoint selection for occlusion handling, combined with next-best-view planning algorithms optimizing information gain per capture.
- Tactile integration: Force/torque sensing for grasp refinement and failure recovery, learning tactile signatures of successful grasps to improve future predictions through haptic feedback loops.
- Reinforcement learning: Learn optimal grasping strategies from interaction, discovering manipulation primitives not specified during training, with sim-to-real transfer accelerating learning through parallelized simulation.
- Digital twin integration: Real-time simulation mirroring physical system for training data augmentation, predictive maintenance, and what-if scenario analysis without disrupting production.
- Multi-object manipulation: Simultaneous multi-part grasping where appropriate, doubling or tripling throughput for compatible component configurations through coordinated multi-gripper systems.

VI. Conclusion

This study proposes an innovative framework that covers the entire pipeline of robotic grasping of flexible automotive trim parts in RGB-D settings. The framework suggests three innovative ideas that work in unison to tackle the inherent difficulties in handling dimensionally varying interlocked objects.

GraspAnnotator Pro's colored point cloud annotation revolutionizes training data creation, achieving 9.4× speedup over traditional methods with 94% error reduction through unified 3D annotation eliminating RGB-to-3D alignment issues. The 96.3% projection accuracy and 100% automatic file pairing

enable rapid, high-quality dataset generation critical for deep learning success. The multi-modal GraspingNet architecture with adaptive radius prediction achieves 94.7% test success rate, demonstrating the effectiveness of early RGB-3D fusion and multi-task learning. Ablation studies confirm 73% position error reduction from fusion and 5.7% success improvement from adaptive radius, validating architectural design choices.

Most notably, the production validation over more than 3,000 cycles resulted in a 93.7% success rate on the first attempt, exceeding the baseline figure of 92.8% while competing with other general-purpose lab equipment in the 88-92% range. Meanwhile, the 42% decrease in cycle time, the 69% decrease in variability, the 62% increase in ergonomics, and the 96.2% system uptime speak to the significant potential for automation in the automotive industry. Quality improvements such as the 94.7% accuracy in defect detection as well as the 87% reduction in customer returns speak to the significant business value provided by the automation solution beyond the direct labor savings.

A. Current Limitations and Scope

The present validation encompasses 10 distinct J-shaped component variants within a single automotive manufacturing facility over 6 weeks of continuous operation. While demonstrating reliability under these conditions, the framework has not been validated across substantially different geometries, materials, or production environments. The 6.3% complete failure rate concentrates primarily in extreme-curvature components (radius $<$ 8mm), representing a known limitation of the current deformation model that requires future investigation. Performance degradation occurs when occlusion exceeds 60%, necessitating multiple retry attempts. Generalization to components beyond the trained distribution remains an open challenge requiring either retraining with expanded datasets or development of few-shot adaptation methods. The annotation process, despite 9.4$\times$ acceleration, still requires 12.2 hours for 2,847 samples, motivating exploration of active learning techniques.

B. Transferability and Generalization to Other Geometries

An important consideration for practical utility is the extent to which the proposed framework transfers to fastener geometries beyond J-shaped clips and to different bin configurations. The system's architecture was deliberately designed with transferability in mind: the GraspAnnotator Pro annotation pipeline is geometry-agnostic, as polygon and suction-circle annotations operate on colored point clouds regardless of object shape. The GraspingNet multi-modal fusion architecture does not embed J-shape-specific priors; the shape specialization resides entirely in the training data. This implies that retraining the model with a new annotated dataset—facilitated by GraspAnnotator Pro—would adapt the system to other clip types (U-clips, C-clips, ring terminals) or general bin-picking scenarios. However, several conditions bound the current generalizability. First, severe non-rigid deformation models specific to J-clips (e.g., the deformation tolerance thresholds) require re-parameterization for geometries with different material stiffness. Second, PLC integration logic for fault tolerance (e.g., retry heuristics) was tuned for the specific target-placement zone configuration and would require re-commissioning in new production layouts. Third, the current pneumatic gripper design is optimized for the contact surface curvature range of J-clips; substantially different contact geometries may require hardware adaptation. Despite these constraints, the modular architecture—separating annotation, perception, planning, and PLC control into distinct layers—provides a practical pathway for rapid adaptation to new part families with manageable engineering effort.

The modularity in the framework's architecture allows it to be adapted to other application areas besides automotive trim parts assembly. Colored point cloud annotation is applicable to other pneumatic grasping situations where precise recognition of the contact surface is required. Similarly, the multimodal deep learning architecture is applicable to other RGB-D manipulation situations. PLC integration as well as the various fault-tolerant control approaches have significant application in the area of industrial automation, though implementation complexity may require specialized expertise in resource-constrained organizations.

C. Specific Technical Challenges and Future Directions

Moving forward, deployment experience reveals three specific technical challenges motivating concrete research directions. First, the concentration of failures under extreme occlusion ($>$60%) indicates need for active vision strategies enabling dynamic viewpoint selection, combined with next-best-view planning algorithms optimizing information gain per capture and physics-based bin manipulation to reduce component interlocking before retry. Second, wire profiles with severe twisting or crushing beyond the deformation model's training distribution constitute 43% of total failures (28 of 65 complete failures), demonstrating limitations that may be addressed through graph neural networks or physics-informed neural networks that explicitly model material properties and geometric constraints for real-time deformation estimation. Third, the annotation burden, despite substantial acceleration, motivates investigation of active learning techniques for selecting maximally informative samples and semi-supervised learning with pseudo-labeling to exploit unannotated production data.

The combination of active vision, tactile sense, and reinforcement learning is expected to address the existing limitations in terms of extreme occlusions and significant deformations. Force/torque sensing for grasp refinement and failure recovery may enable learning tactile signatures of successful grasps to improve future predictions through haptic feedback loops. Reinforcement learning approaches may discover optimal grasping strategies from interaction, with manipulation primitives not specified during training, leveraging sim-to-real transfer to accelerate learning through parallelized simulation. Few-shot learning, as well as the transfer from simulated to realistic scenarios, is expected to facilitate quick deployments across various part geometries with minimal further training, utilizing domain-randomized synthetic data created from parametric CAD models.

Moving forward from the optimization of specific components towards the development of more comprehensive human-centric automation techniques as presented in the work above presents the future direction towards Industry 5.0—a more

resilient, as well as more sustainable form of manufacturing, where intelligent automation techniques complement human capabilities. The combination of specific annotation tools, adaptive deep learning techniques, as well as robust control techniques as presented in the work above presents a future direction towards the development of advanced automation techniques in the field of manufacturing. Thus, the 62% improvement in terms of ergonomics, as well as the 93.7% reliability, as presented in the work above presents the future capabilities in terms of the quality, as well as the health of the employees—a synergistic outcome required in the future of manufacturing where advanced AI and robotics simultaneously improve quality, throughput, and worker health.

REFERENCES

[1] N. Van Nguyen et al., "BOP challenge 2024 on model-based and model-free 6D object pose estimation," arXiv:2506.00599, 2025.

[2] X. Li et al., "A sim-to-real object recognition and localization framework for industrial robotic bin picking," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 3961-3968, 2022.

[3] K. Kleeberger and M. F. Huber, "Single shot 6D object pose estimation," in *Proc. IEEE ICRA*, pp. 6239-6245, 2020.

[4] B. Russell et al., "LabelMe: A database and web-based tool for image annotation," *International Journal of Computer Vision*, vol. 77, pp. 157-173, 2008.

[5] Z. Zhang, "A flexible new technique for camera calibration," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 11, pp. 1330-1334, 2000.

[6] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "PointNet++: Deep hierarchical feature learning on point sets in a metric space," in *Proc. NeurIPS*, 2017.

[7] J. Mahler et al., "Dex-Net 2.0: Deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics," in *Proc. RSS*, 2017.

[8] D. Morrison, P. Corke, and J. Leitner, "Learning robust, real-time, reactive robotic grasping," *International Journal of Robotics Research*, vol. 39, no. 2-3, pp. 183-201, 2020.

[9] A. Zeng et al., "Robotic pick-and-place of novel objects in clutter with multi-affordance grasping and cross-domain image matching," in *Proc. IEEE ICRA*, 2018.

[10] C. Wang et al., "DenseFusion: 6D object pose estimation by iterative dense fusion," in *Proc. IEEE CVPR*, pp. 3343-3352, 2019.

[11] A. Sanchez-Gonzalez et al., "Learning to simulate complex physics with graph networks," in *Proc. ICML*, 2020.

[12] B. Settles, "Active learning literature survey," Computer Sciences Technical Report 1648, University of Wisconsin-Madison, 2009.

[13] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proc. ICML*, 2017.

[14] J. Tobin et al., "Domain randomization for transferring deep neural networks from simulation to the real world," in *Proc. IEEE/RSJ IROS*, 2017.

[15] B. Sekachev et al., "Computer Vision Annotation Tool (CVAT)," https://github.com/opencv/cvat, 2019.

[16] M. Sundermeyer et al., "Contact-GraspNet: Efficient 6-DoF grasp generation in cluttered scenes," in *Proc. IEEE ICRA*, pp. 13438–13444, 2021.

[17] H. Fang et al., "AnyGrasp: Robust and efficient grasp perception in spatial and temporal domains," *IEEE Transactions on Robotics*, vol. 39, no. 5, pp. 3929–3945, 2023.

[18] B. Wen et al., "FoundationPose: Unified 6D pose estimation and tracking of novel objects," in *Proc. IEEE CVPR*, pp. 17868–17879, 2024.

[19] J. Redmon et al., "You only look once: Unified, real-time object detection," in *Proc. IEEE CVPR*, pp. 779–788, 2016.

[20] G. Jocher et al., "Ultralytics YOLOv8," https://github.com/ultralytics/ultralytics, 2023.

[21] A. Ichnowski et al., "Dex-NeRF: Using a neural radiance field to grasp transparent objects," in *Proc. CoRL*, 2022.

[22] H. Fang et al., "GraspNet-1Billion: A large-scale benchmark for general object grasping," in *Proc. IEEE CVPR*, pp. 11444–11453, 2020.

[23] "Labelbox: The leading training data platform for data labeling," https://labelbox.com, 2023.