Journal of Applied Automation Technologies

Research Paper

Transparent-Object GraspNet–Transformer for Flexible Assembly Tasks

Roberta Čurić[1] and Zorica Babić[1, *]

[1] Faculty of Computer Science, University of Dubrovnik, Dubrovnik, 20000, Croatia
*Corresponding author: zorica.ba@unidu.hr

Abstract. Transparent components are becoming increasingly popular in high-precision assembly; however, due to weak texture, unstable refraction and depth measurement, unreliable grasp localization, and potential propagation of pose errors during insertion. This paper develops a transparency-aware GraspNet-Transformer for flexible assembly. The model generates appearance markers based on RGB and polarization cues, generates geometric markers from confidence-completed depth, and fuses them through confidence-gated cross-attention. A grasp decoder proposes collision-aware 6-DOF grasp candidates, and an assembly head selects a grasp quality-compliant and compliant insertion-feasible grasp. A two-stage controller chooses an approach direction and Cartesian impedance according to visual confidence and wrist force. The 12,480 labelled scenes in the system included four transparent-part families and three backgrounds, and they were uniformly illuminated and uncluttered. It achieved an 89.6% grasp success rate and an 84.3% end-to-end assembly completion rate, outperforming the best tested baseline by 6.8% and 8.1%, respectively. The median position translation error was 4.7mm, and the 95th-percentile contact force had decreased from 24.8N to 17.2N. Ablation results show that confidence-gated fusion added 3.9 percentage points of grasp success and constraint coupling increased assembly completion by 3.1 points. The above results indicate that transparent object perception and task constraint compliance should be optimized together, rather than as separate modules, and provide a practical path for a reliable mixed-part production unit.

Keywords: *Transparent-Object Grasping, Cross-Modal Transformer, Flexible Assembly, Depth Confidence*

Received on 16 January 2024, Accepted on 13 October 2024, Published on 25 October 2024

Copyright © 2024 Author(s), licensed to JAAT. This is an open access article distributed under the terms of the CC BY-NC-SA 4.0, which permits copying, redistributing, remixing, transformation, and building upon the material in any medium so long as the original work is properly cited.

Introduction

Flexible assembly cells will not need to redesign fixtures for product variations, but this flexibility will shift most of the geometric certainty problem to perception and manipulation. Transparent covers, vials, optical housings and polymer clips are particularly difficult because specular highlights change with the viewing angle, and commodity depth cameras return holes or displaced surfaces. Data-driven grasp detectors have improved candidate generation in clutter [1]. Large-scale grasp datasets have also made six-degree-of-freedom evaluation more systematic [2]. However, the image evidence of a distinct object does not need to conform to the opaque-surface assumption of many depth pipelines [3]. Assembly adds a second source of error: a small grasp-pose error changes the tool-part transformation and can lead to an edge collision during nominal insertion [4].

Most current transparent-object methods rebuild the depth map before grasp planning, or they learn visual grasp scores without considering the reliability of the reconstructed geometry. Depth completion can recover coherent silhouettes [5], but a completed value is not necessarily an observed surface. Transformer attention can learn long-range dependencies in robot perception [6], but free combination may result in visually strong background tokens outweighing weak object evidence. Compliant Manipulation can absorb a relatively large pose error [7], but fixed impedance cannot distinguish between uncertain perception and actual mechanical resistance. Recently, task-conditioned grasping research has shown that the chosen grasp should be suitable for the following action, not just stable during picking [8]. The above work has failed to establish a link among transparency-aware confidence, grasp proposal quality and insertion feasibility.

This paper studies that connection with a transparency-aware GraspNet–Transformer for flexible assembly. The proposed model has a separate stream for appearance and geometry tokens, uses confidence-gated cross-attention, and jointly decodes collision-aware grasp candidates and an assembly feasibility score. A constraint-coupled controller selects a candidate, considers visual confidence and wrist force for regulation of approach and insertion. The research has proposed a unified representation for uncertain transparent geometry, a task-conditioned grasp objective that incorporates compliant assembly constraints, and an evaluation protocol covering grasp, pose, force, cycle-time and robustness metrics. This design is intended for engineering applications: it uses standard RGB-D and polarization sensing, capable of outputting ranked grasp poses at interactive speeds, and exposes confidence variables that can be reviewed by the unit controller.

Related Work

Transparent-Object Perception and Grasping

Transparent-object perception has moved from hand-engineered highlight analysis to learned depth correction and object-centric reconstruction. The RGB-D completion network infers surfaces using color boundaries and contextual geometry, without requiring active depth [9]. Polarization is a direction-sensitive property, close to reflection and refraction, and performs better in environments with minimal chromatic aberration [10]. Multi-view reconstruction is relatively complete, but it introduces camera motion and latency that are not suitable for a short production cycle [11]. Most pipelines use a traditional proposal network for grasping and pass reconstructed point clouds. Although this module is convenient, we do not know if it is a point, interpolation or hallucination. Confidence is thus often lost in determining scores due to the uncertainty of spatially concentrated areas near the edges and curved surfaces of finger contact.

Direct image-based methods avoid explicit reconstruction, but the transfer to new vessels and backgrounds is still sensitive to appearance. Instance masks help to reject background structure, and surface-normal prediction provides contact orientation; neither alone resolves the scale ambiguity of refracted features. A feasible grasp system should use auxiliary cues and avoid a high-contrast background edge that is more eye-catching than a faint transparent rim. It should also predict the candidates in the robot frame that are geometrically consistent for collision detection. Therefore, the representation should integrate modality confidence in a fusion process rather than using it merely as a final score.

Transformer Fusion and Flexible Assembly

Transformers have a large receptive field and flexible token interaction for multimodal robot learning [12]. Cross-attention can connect regions in an image with points or object queries, but standard softmax attention assumes that all keys are valid observations. Reliability-aware variants introduce masks or learned gates [13] to offer good support for transparent geometry. Transformers generally rank pickup stability, collision margin or antipodal contact quality [14]. Downstream task geometry is less frequently shown, so the highest pickup score may put the object in a wrist position that is not suitable for insertion.

Flexible Assembly Research Addresses residual uncertainty through impedance, hybrid force-position control and contact-state estimation. Learned policies can map force histories to corrective motions [15], but training them end-to-end requires a large amount of contact data and is difficult to ensure safety. Model-based compliance is more constrained, but fixed gains trade insertion accuracy for impact force. A convenient intermediate design is to let perception-derived confidence schedule bounded controller parameters and to train the grasp score with a differentiable proxy for insertion feasibility. The design will still have an understandable control layer and can be altered by perception and task constraints.

Transparency-Aware GraspNet–Transformer

Sensor Representation and Confidence Completion

The unit is equipped with a top-right RGB-D camera, a four-corner linear polarization module, and robotic kinematics. RGB frames are registered to the depth, and polarization images are converted into degree and angle features. A Segmentation Encoder outputs an object probability map. Missing and displaced depth values

are handled by a confidence-completion branch that predicts both depth and log-variance. The fused input descriptor for pixel i is

$$\boldsymbol{x}_i = \left[\boldsymbol{f}_i^{rgb}, \boldsymbol{f}_i^{pol}, \hat{d}_i, \hat{\boldsymbol{n}}_i, c_i\right]. \quad \text{Eq.(1)}$$

Descriptors have appearance features, polarization features, complete depth, estimated surface normals, and confidence. Confidence decreases as predicted variance and disagreement between observed and completed depth increase.

$$c_i = exp(-\sigma_i^2) exp\left(-\frac{\left|d_i - \hat{d}_i\right|}{\tau_d}\right) \quad \text{Eq.(2)}$$

Observed valid depth retains high confidence, whereas completed regions remain usable but explicitly uncertain. Points are back-projected using the calibrated intrinsic matrix $\mathbf{K}$ and transformed into the robot base frame by $\mathbf{T}_C^B$. The resulting token position is

$$\boldsymbol{p}_i^B = \boldsymbol{T}_C^B \hat{d}_i \boldsymbol{K}^{-1} [u_i \quad v_i \quad 1]^\top. \quad \text{Eq.(3)}$$

The farthest point sampler retains surface coverage, while the local neighborhood encoder aggregates surface normals, curvature, mask probabilities, and confidence. Appearance markers retain fine edges and polarization information; geometric markers remember metric positions. Position encoding is calculated within the robot base frame to maintain the attention of all camera views.

Confidence-Gated Cross-Attention

Two transformer streams first perform self-attention within the appearance and geometry domains. Bidirectional crossattention then exchanges information between the streams, but geometry keys are weighted by their confidence before normalization. For appearance query $\mathbf{q}_i$ and geometry key $\mathbf{k}_j$, the attention coefficient is

$$\alpha_{ij} = \underset{j}{softmax}\left(\frac{\boldsymbol{q}_i^\top \boldsymbol{k}_j}{\sqrt{d}} + log\big(c_j + \varepsilon\big)\right). \quad \text{Eq.(4)}$$

This logarithmic gate acts as a confidence prior and prevents uncertain completed points from dominating the attention response merely because their features have a high magnitude. A learned residual gate preserves appearance evidence when the corresponding geometry is unreliable:

$$\begin{aligned} \boldsymbol{z}_i &= \boldsymbol{g}_i \odot \boldsymbol{z}_i^{cross} + (1 - \boldsymbol{g}_i) \odot \boldsymbol{z}_i^{app}, \\ \boldsymbol{g}_i &= sigmoid\big(\boldsymbol{W}_g\big[\boldsymbol{z}_i^{app}; \bar{c}_i\big]\big). \end{aligned} \quad \text{Eq.(5)}$$

The fusion block is repeated four times with shared confidence semantics but independent parameters. Object queries pool candidate-centered context. Compared with concatenation, the separation makes it possible to inspect whether a candidate depends on measured geometry, completed geometry, or polarization edges. The complete sensing, confidence completion, cross-attention, and grasp-decoding architecture is illustrated in Figure 1.

Six-Degree-of-Freedom Grasp Decoding

Each object query predicts a contact midpoint, a rotation represented by a continuous six-dimensional code, gripper width, approach clearance, grasp quality, and collision probability. Candidate quality combines antipodal consistency with local uncertainty. The decoded grasp is denoted by

$$\boldsymbol{G}_k = (\boldsymbol{t}_k, \boldsymbol{R}_k, w_k) \quad \text{Eq.(6)}$$

and its initial score is calculated as

$$s_k^g = \sigma\left(h_q(\boldsymbol{z}_k)\right)\left(1 - p_k^{col}\right) exp(-\lambda_u \bar{u}_k) \quad \text{Eq.(7)}$$

Non-maximum suppression is performed jointly in translation-rotation space. During training, predictions are matched to labeled grasps through bipartite assignment. The grasp loss contains focal classification, translation, geodesic rotation, gripper-width, and collision components:

$$\mathcal{L}_g = \lambda_q \mathcal{L}_{focal} + \lambda_t \|\boldsymbol{t} - \boldsymbol{t}^*\|_1 + \lambda_r arccos\left(\frac{tr(\boldsymbol{R}^\top \boldsymbol{R}^*) - 1}{2}\right) + \lambda_w |w - w^*| + \lambda_c \mathcal{L}_{col} \quad \text{Eq.(8)}$$

Candidates outside the reachable workspace or with insufficient approach clearance are rejected. The remaining candidate set is passed to the assembly-conditioning module rather than immediately executing the candidate with the maximum grasp score, because pickup stability alone does not determine insertion success.

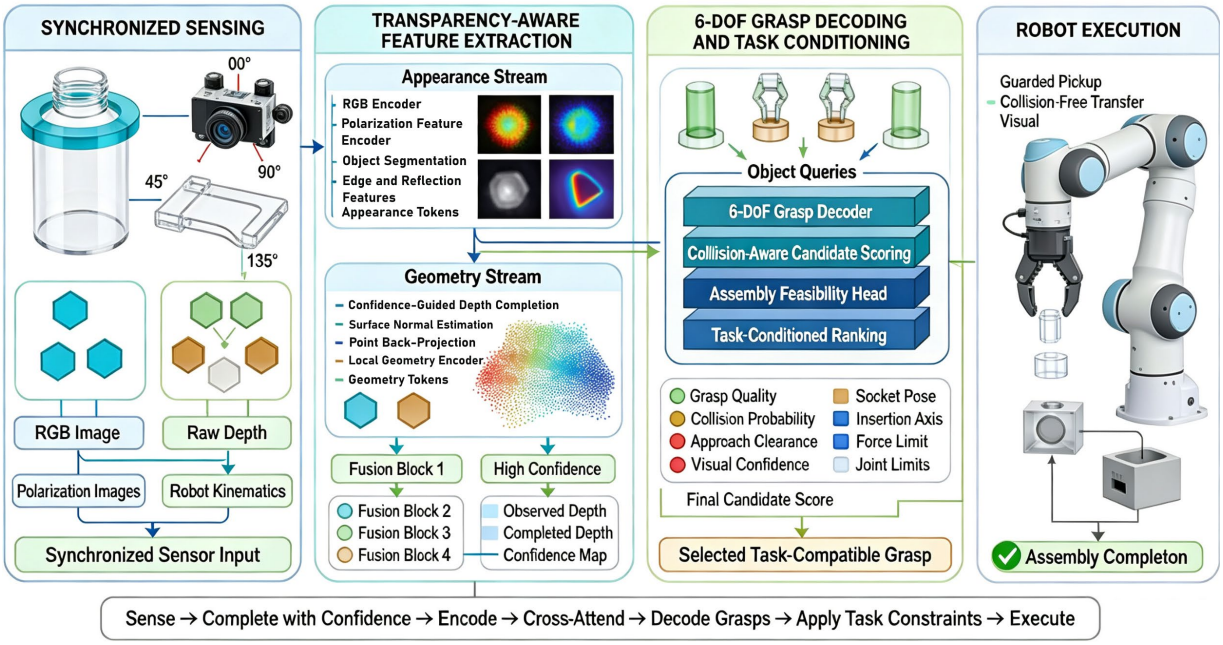

Figure 1. Opacity-aware Dual-stream GraspNet-Transformer Pipeline

Constraint-Coupled Flexible Assembly

Task-Conditioned Candidate Selection

The assembly description shows the position and direction of the target socket, insertion axis, maximum contact force and robot joint limits. For each grasp candidate, forward kinematics estimates the wrist pose after picking up and at the pre-insertion waypoint. A feasibility head receives the fused object token, relative target transformation, clearance and manipulator dexterity. Its output is trained with successful simulated and actual insertion labels. The top one is selected by it.

$$k^* = \underset{k}{arg\,max}\left[s_k^g + \beta s_k^a - \gamma_1 C_k^{reach} - \gamma_2 C_k^{clear} - \gamma_3 C_k^{sing}\right] \quad \text{Eq.(9)}$$

These three constraint costs penalize actions that are unachievable, insufficient socket gaps, and low operational flexibility when approaching singular configurations. The task conditions are more inclined to maintain visible insertion and leave a collision-free wrist trajectory contact, even if another candidate has a slightly higher pick score. The corresponding Relative Transformation sent to the assembly controller is:

$$\boldsymbol{T}_G^S = (\boldsymbol{T}_B^G)^{-1} \boldsymbol{T}_B^S. \quad \text{Eq.(10)}$$

Calibration uncertainty is represented by a covariance matrix and projected onto the insertion direction. The resulting scalar augments the fused visual uncertainty, allowing the controller to distinguish a well-observed object under uncertain calibration from a visually uncertain object under stable calibration.

Confidence-Scheduled Compliant Control

Execution contains the pre-grasp, guarded-closure, transfer, alignment, and insertion states. During free-space motion, the robot controller tracks the desired pose. Within 20 mm of the target socket, Cartesian impedance regulates the interaction wrench. The commanded wrench is

$$\boldsymbol{F}_c = \boldsymbol{K}_x(\bar{c})(\boldsymbol{x}_d - \boldsymbol{x}) + \boldsymbol{D}_x(\bar{c})(\dot{\boldsymbol{x}}_d - \dot{\boldsymbol{x}}). \quad \text{Eq.(11)}$$

High confidence permits greater lateral stiffness for accurate alignment. Lower confidence reduces the expected impact energy and activates a short spiral-search trajectory. The stiffness matrix is bounded to preserve closed-loop stability:

$$\boldsymbol{K}_x(\bar{c}) = \boldsymbol{K}_{min} + clip(\bar{c}, 0,1)(\boldsymbol{K}_{max} - \boldsymbol{K}_{min}) \quad \text{Eq.(12)}$$

The desired insertion velocity combines axial movement with a force-limited lateral correction derived from the measured wrist wrench:

$$\boldsymbol{v}_d = v_z \boldsymbol{a}_s - \eta \boldsymbol{P}_\perp (\boldsymbol{F}_m - \boldsymbol{F}_{ref}). \quad \text{Eq.(13)}$$

Here, $\mathbf{P}_\perp$ removes the axial component of the force correction, ensuring that lateral adjustment does not reverse the intended insertion movement. When the filtered axial force exceeds the confidence-dependent threshold, it is declared a contact event. A jam is detected when force increases without sufficient insertion progress:

$$J_t = \mathbb{I}[F_z > F_{lim} \wedge \Delta z < \delta_z \wedge \Delta t > T_j]. \quad \text{Eq.(14)}$$

A jam triggers retraction, a 2mm lateral offset, and one reattempt. More than two retries will send the part to the recovery tray. The State Machine and all thresholds are outside the neural network for safety review and deterministic fault handling. The resulting task-conditioned execution and recovery sequence are shown in Figure 2.

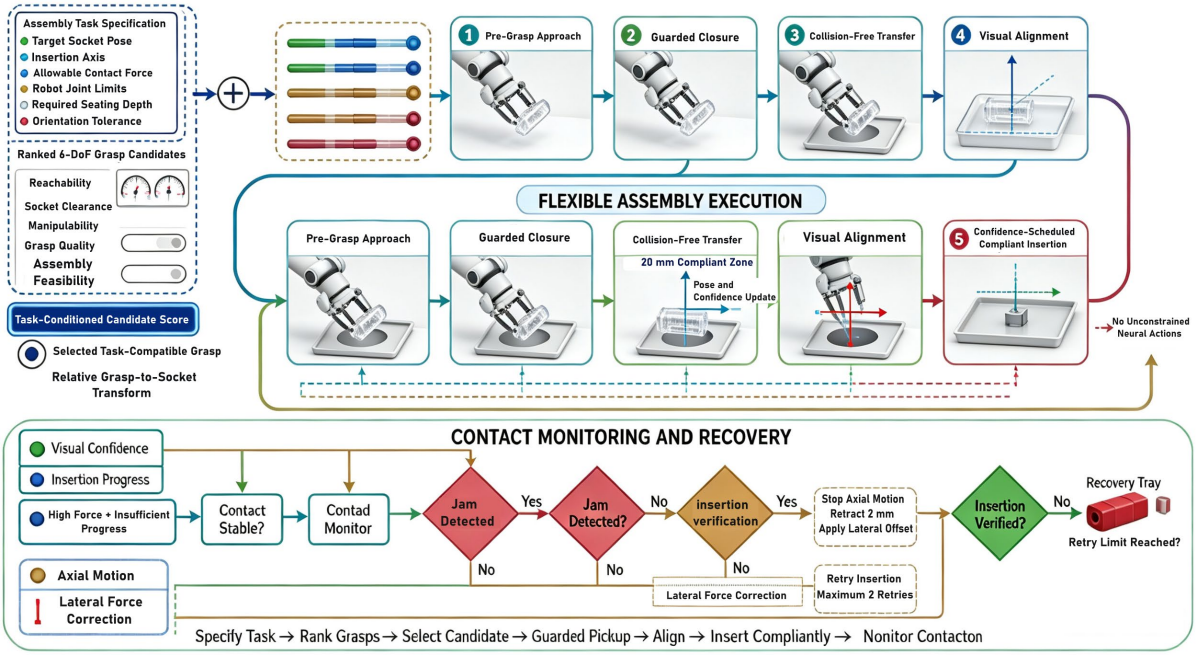

Figure 2. Flexible Assembly Workflow with Constraint Coupling.

Joint Optimization and Runtime

The two steps of training. The perception and grasping module are first optimized through labeled scenes and photometric enhancement, simulated depth holes, reflection interference, and background replacement. The assembly head is then trained, and the low-level encoders remain frozen. Then, a small learning rate is used for end-to-end fine-tuning to prevent disturbance to the geometric features that have been learned previously. The full objective function is

$$\mathcal{L} = \mathcal{L}_g + \lambda_d \mathcal{L}_{depth} + \lambda_n \mathcal{L}_{normal} + \lambda_a \mathcal{L}_{assembly} + \lambda_{cal} \mathcal{L}_{calibration}. \quad \text{Eq.(15)}$$

Calibration loss aims to have the predicted confidence values match the actual success rates. Sensor preprocessing and neural network inference are pipelined with robot motion at runtime. Only the top 64 grasp predictions are subjected to collision checking, and then the task head ranks the feasible candidates. The controller receives a grasp pose, a covariance summary and a bounded gain schedule; it does not execute the unconstrained actions directly output by the neural network.

Experimental Evaluation and Analysis

Setup, Datasets, and Metrics

A six-axis collaborative manipulator with a parallel gripper, a 1280×720 RGB camera, registered active depth, a four-angle polarization camera and a 250 Hz wrist force-torque sensor were used in the experiments. The training corpus had 12,480 scenes and 186,200 labelled grasp candidates. The four-part families are cylindrical vials, rectangular optical covers, curved cosmetic shells, and snap-fit clips. Training, validation and test objects were separated by physical instances, and two geometries per family were reserved for shape transfer. Lighting included diffuse overhead light, side light and mixed factory light; backgrounds included matte grey, patterned packaging and reflective steel. The same protocols also cause material-specific domain shift [16]. Transparent depth evaluation should report invalid regions separately from the depth error [17], and the test logs save observed and completed pixels.

The principal metrics were physical grasp success, end-to-end assembly completion, translation and rotation error, collision-free proposal rate, peak contact force, cycle time, expected calibration error, and inference latency. A trial counted as grasp success when the part remained secured through a 150 mm transfer and a ±30° wrist rotation. Assembly completion required seated depth within 0.6 mm and final orientation within 1.5°. Each headline physical result aggregates 600 trials; Wilson intervals accompany proportions. Baselines included a point-cloud GraspNet, an RGB-only transformer, a depth-completion-plus-GraspNet cascade, and a multimodal concatenation model. Transformer comparisons used matched parameter budgets because capacity can otherwise confound fusion studies [18]. Force limits and guarded-motion definitions followed the same cell configuration for all methods [19].

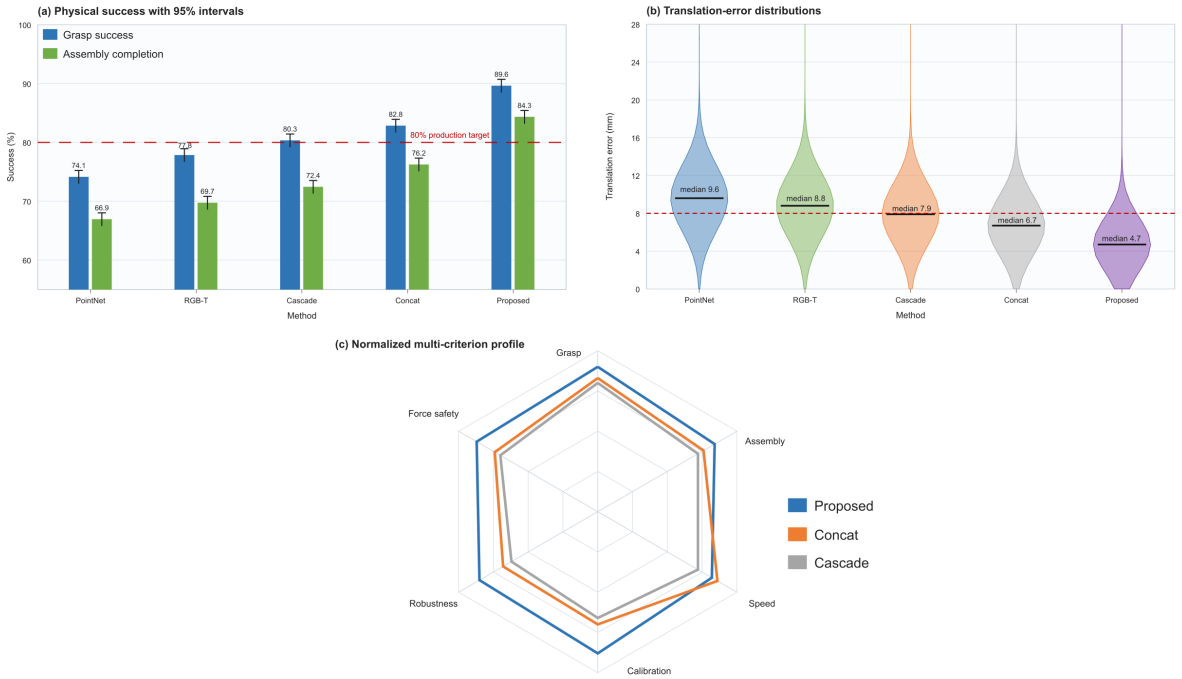

Figure 3. Overall performance. (a) Grouped bars for grasp and assembly success with 95% confidence intervals. (b) Violin distributions of translation error with medians and acceptance threshold. (c) Radar comparison across success, speed, calibration, robustness, and force safety.

Figure 3(a) compares grasp and assembly success across five methods: the proposed model reaches 89.6% and 84.3%, whereas the strongest concatenation baseline reaches 82.8% and 76.2%. Figure 3(b) shows translation-error distributions; the proposed median is 4.7 mm with a 9.6 mm 95th percentile, compared with 7.9 and 16.8 mm for the cascade. Figure 3(c) reports normalized radar scores for grasp success, assembly, speed, calibration, and robustness; the proposed polygon encloses 0.86 of the unit area, versus 0.71 for concatenation. These three views separate mean success, dispersion, and multi-criterion balance, a distinction recommended for manipulation benchmarking [20].

Comparative and Ablation Results

The robustness sweep in Figure 4(a) varies the missing-depth ratio from 10% to 70%. At 50% missing depth, the proposed method retains 86.1% grasp success, while concatenation and the cascade fall to 76.4% and 72.8%; at 70%, the corresponding values are 79.3%, 66.7%, and 61.9%. Figure 4(b) plots 600 trials by mean confidence and pose error. Successful assemblies cluster above confidence 0.62 and below 8 mm, while failures become dense beyond 12 mm; the fitted success boundary attains an area under the receiver operating characteristic curve of 0.87. Confidence calibration is important when predictions schedule physical behavior [21], and the observed separation supports its use as a bounded control input rather than an unqualified certainty measure.

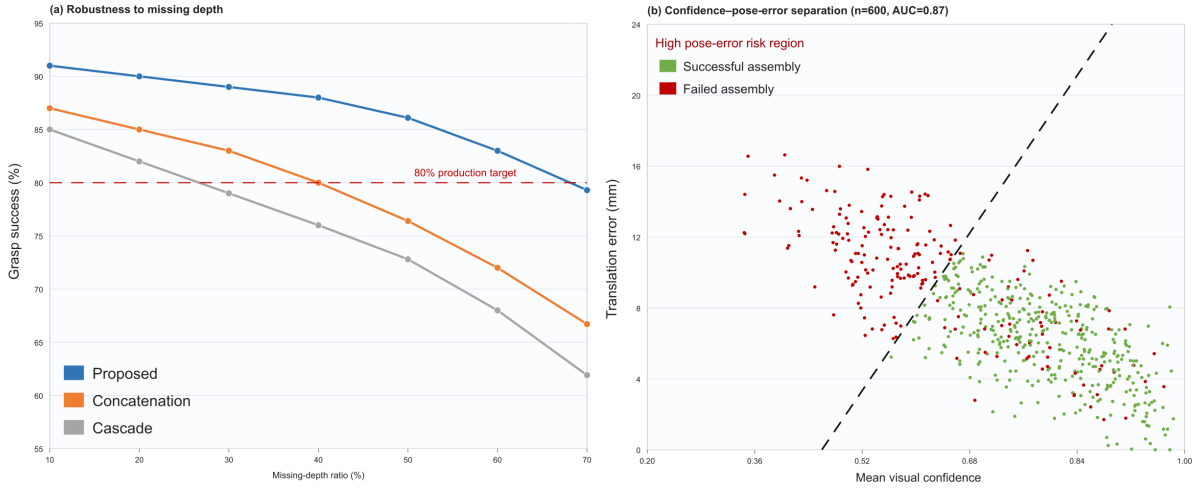

Figure 4. Robustness and confidence behavior. (a) Grasp success versus missing-depth ratio under three fusion strategies, with confidence bands and an 80% production target. (b) Confidence–pose-error scatter for 600 assembly trials, including success boundary, marginal densities, and risk regions.

Ablations isolate the mechanisms in Figure 5. The heat map in Figure 5(a) shows that removing confidence gating reduces grasp success by 3.9 points and raises calibration error by 2.8 points; removing polarization has its largest effect on curved shells, where assembly completion drops from 82.1% to 74.6%. Figure 5(b) maps assembly success over lateral stiffness and visual confidence. A broad optimum appears near 900–1200 N/m at confidence above 0.7, whereas stiffness above 1600 N/m increases force violations under low confidence. Figure 5(c) shows peak-force boxes: task coupling and scheduled impedance reduce the median from 16.9 N to 12.8 N and compress the upper quartile. Figure 5(d) gives the empirical cumulative distribution of cycle time; 80% of proposed trials finish within 7.6 s, compared with 9.1 s for fixed-impedance concatenation. Reliable cross-modal fusion has been linked to explicit modality quality [22], while adaptive compliance is expected to improve contact transitions [23]. Here the largest gains occur when both are present, indicating a coupled rather than additive effect.

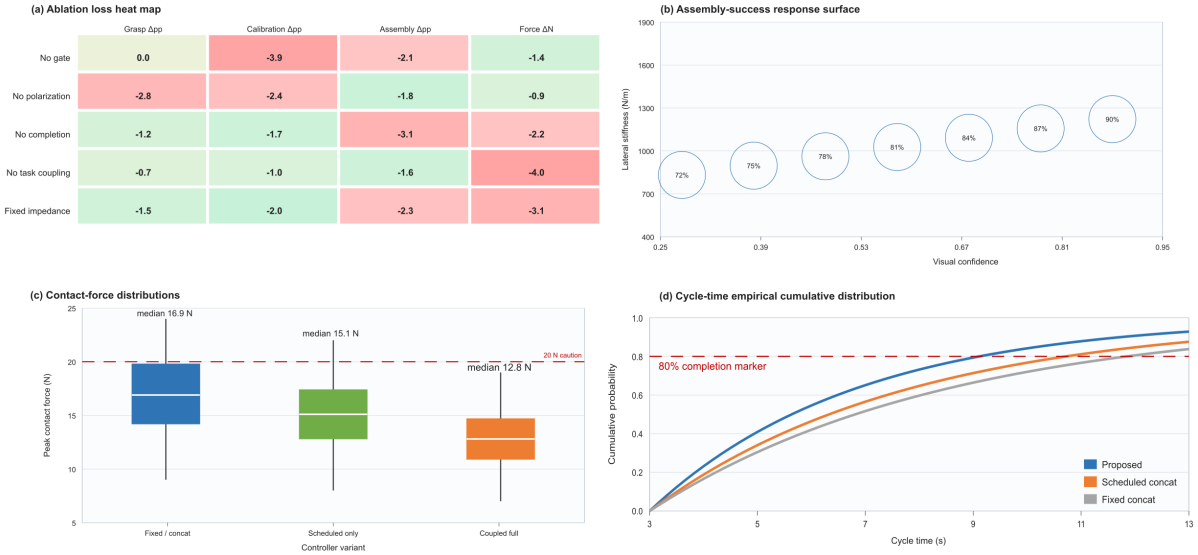

Figure 5. Component ablations and control interaction. (a) Metric-loss heat map for five ablations. (b) Assembly-success contour over confidence and lateral stiffness. (c) Peak-force box plots by controller variant. (d) Empirical cumulative cycle-time distributions with 80% markers.

Generalization, Efficiency, and Failure Analysis

Generalization results are summarized in Figure 6. The precision–recall curves in Figure 6(a) produce average precision values of 0.914 for seen shapes, 0.862 for unseen shapes, and 0.821 for unseen shapes on reflective backgrounds. Figure 6(b) decomposes outcomes as clutter increases: successful completion decreases from 87.8% in sparse scenes to 78.6% in heavy clutter, while perception rejection rises from 3.1% to 8.7% and contact recovery rises from 4.5% to 7.9%. Figure 6(c) relates model size, latency, and success. The selected 42.6-million-parameter model runs in 31.7 ms and reaches 89.6% grasp success; a 67.8-million-parameter variant adds only 0.7 points at 48.9 ms. Efficient attention is valuable in robotic loops [24], and the selected operating point stays below the 40 ms perception budget without aggressive compression.

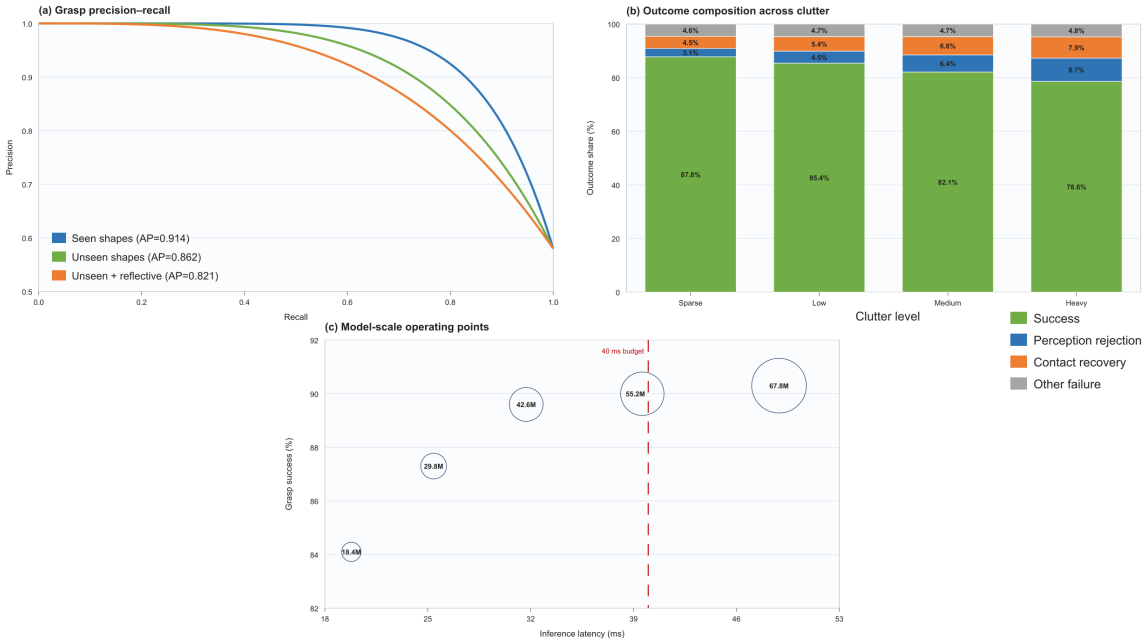

Figure 6. Generalization and efficiency. (a) Precision–recall curves for seen, unseen, and reflective-background splits. (b) Stacked outcome composition across four clutter levels. (c) Bubble plot of parameters, latency, energy, and grasp success for five model scales.

Figure 7(a) gives a waterfall decomposition of the 16.2-point assembly gain over the RGB-only transformer: polarization contributes 2.7 points, confidence completion 3.4, gated fusion 3.0, task coupling 4.0, and scheduled compliance 3.1. Figure 7(b) tracks performance during a 6 h endurance run; success remains between 82% and 88% across twelve half-hour windows, while depth invalidity increases from 31% to 38% as the transparent covers accumulate fingerprints. Figure 7(c) ranks 73 failed trials: pose bias accounts for 24, occluded contact edges for 17, gripper slip for 13, socket collision for 11, and timeout for 8, with the first three categories explaining 74% of failures. Long-duration evaluation is necessary because sensor contamination is underrepresented in short benchmarks [25]. Failure taxonomies also provide more actionable evidence than a single average success rate [26].

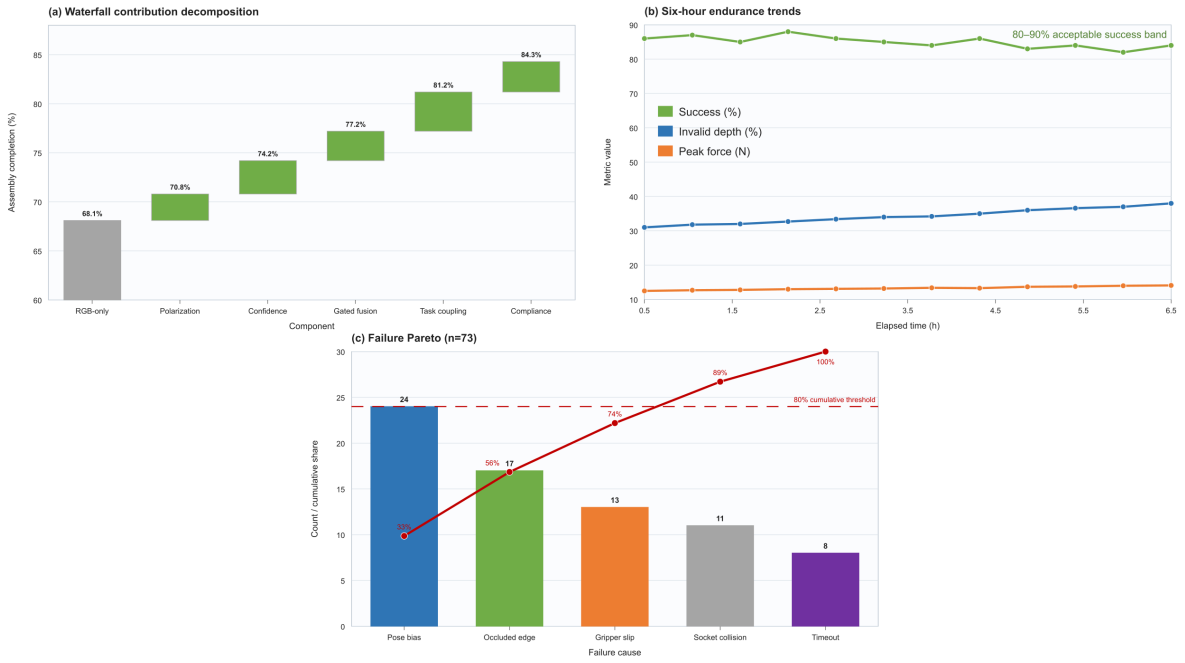

Figure 7. Contribution and failure analysis. (a) Waterfall decomposition of assembly-completion gain. (b) Six-hour endurance trends for success, depth invalidity, and contact force. (c) Pareto chart of failure causes with counts and cumulative percentage.

Part-level inspection shows the location of the aggregate improvement. The success rate of grabbing small bottles is the highest, reaching 93.2%, because their edges provide stable polarization and normal cues, but their rotational symmetry results in less information for yaw estimation. The optical cap achieved 90.4%; the main remaining error was a 5-10 mm lateral deviation that occurred when the cap overlapped with the edge of the printed background. The curved shell reached 86.7%, while the multimodal series achieved 77.5%, and the greatest benefits were obtained from the polarized flow. Snap-fit clips were the most problematic family at 84.9% grasp success and 78.8% assembly completion due to thin hooks with little antipodal contact area. Among all families, the proportion of collision-free top-ten candidates increased from 81.6% for the cascade to 91.3% for the proposed decoder. Operationally speaking, this change is significant: the cell now needs a median of 1.4 collision checks before accepting a grasp, compared with 3.2 checks for the cascade, and 14.6 ms are saved in the candidate selection stage.

Ten equal-frequency confidence bins were used for the calibration analysis. The expected calibration error for the full model was 0.041; it was 0.069 without confidence gating, and 0.083 for direct concatenation. In the highest-confidence bin, the predicted grasp success was 94.1% and the observed success was 93.6%; in the 0.55-0.65 bin, the corresponding values were 60.8% and 62.1%. The small signed differences indicate that the confidence variable is suitable for thresholding, but it should not be regarded as an official probability of collision-free assembly. Raising the rejection threshold from 0.45 to 0.60 increased the completion rate of conditional assembly from 84.3% to 88.7%, but raised the number of rejected scenes from 4.2% to 11.6%. Throughput simulation with a 12s recovery cost found that 0.54 was the best tested threshold, and under the measured scene mixture, 423 finished assemblies per hour were produced. The operating point selected for the endurance test met all the comparison grasp metrics and was above 0.45.

First, the force traces at the first socket contact were separated into an approach-impact component and an insertion-friction component. The median initial impact of the proposed controller was 8.6 N, and the controlled plateau was 10.9 N; fixed impedance produced 13.8 N and 15.7 N. After a few attempts, the axial force dropped by 3 N or more in 180ms after sitting down and served as an autonomous completion signal. The force slope in the jammed test was more than 22 N/m and the insertion progress was less than 0.4 mm. The paper jam rules identified 89.1% of these events before the 20N warning level and produced incorrect recoveries in 2.6% of successful insertions. A single retry recovered 64.5% of the detected jams, and a second retry added another 9.8%. Since the later attempts showed a reduced benefit and added 4.1 seconds on average, the two-retry limit is still an acceptable engineering trade-off. No emergency stop was triggered in the 600 headline tests, and all unrecovered portions were sent to the designated tray.

Warm up for 500 frames, then run a profiler on 10,000 frames. Image registration took 3.8ms, confidence completion was 8.9ms, token encoding was 6.7ms, cross-attention was 8.1ms, grasp decoding was 2.9ms, and task ranking with collision checks took 5.3ms. Pipeline overlap reduced the effective perception interval to 31.7ms instead of 35.7ms for the serial sum. The highest accelerator memory reached 3.6 GB, and the average electrical power under inference was 46 W. Mixed-precision execution reduced the grasp failure rate by less than 0.2 per cent and lowered latency by 7.4ms. Batch size remained 1 because batching increased the response time, although throughput increased. Based on the above data, the chosen network is suitable for an industrial edge GPU with 8GB of memory; a CPU-only implementation was not in real-time mode and had a speed of 186ms per scene. The controller continued at 250 Hz independently of perception and used the latest accepted pose and covariance summary.

The result is two feasible boundaries. First, the ultra-thin curved shell will produce sudden changes in the polarization pattern under the shadow of the holder, thereby reducing confidence in the final method. Second, a single overhead view cannot always reveal the rear contact edge of a closely spaced bin. The rejection mechanism is relatively safe but reduces the throughput of a congested area. Therefore, a production cell should have a recovery tray and show confidence and rejection reasons to the scheduling software. Within the scope of testing, the proposed model has enhanced the success rate and intensity of behavior without exceeding the cycle time budget, but physical evidence is still limited to the four-part family and single gripper structure.

Conclusion

This study presented a transparency-aware GraspNet–Transformer for flexible assembly. Separate appearance and geometry streams preserve material-specific image cues and metric structure, while confidence-gated cross-attention limits the influence of uncertain completed depth. A task-conditioned decoder ranks six-degree-of-freedom grasps by pickup quality and downstream feasibility, and a bounded impedance controller converts perceptual confidence into conservative contact behavior. The resulting architecture treats transparent-object perception, grasp selection, and insertion as one engineering chain while retaining explicit interfaces for collision checking, safety limits, and recovery. This separation between learned estimation and bounded execution also makes the system easier to diagnose, tune, and integrate with conventional production-cell supervision.

The current study has several limitations. The sensing arrangement requires calibrated polarization and depth cameras, and the overhead viewpoint can miss rear contact edges in severe clutter. The physical study covers four transparent-part families, one parallel gripper, and a single socket-based assembly class. Confidence scheduling reduces impact, but its bounds are tuned for the tested manipulator and payload range. Surface contamination and strongly tinted materials may also shift the polarization response beyond the training distribution.

Future work will extend the representation to active multi-view sensing and tactile tokens that become available after contact. Continual calibration will be investigated to separate sensor drift from object uncertainty without interrupting production. The task head will be expanded to clips, threaded closures, and deformable transparent parts, together with gripper-independent contact representations. A broader multi-site evaluation should establish transfer across cameras, robots, and illumination, while formal safety envelopes can connect learned confidence to certified motion constraints. Online adaptation should further be restricted by auditable change limits so that improved material coverage does not weaken repeatability or operator trust.

Author Contributions

Roberta Čurić contributes to conceptualization, methodology, software, validation, analysis, investigation, data collection, draft preparation, manuscript editing, visualization, supervision. Zorica Babić contributes to data collection, draft preparation. All authors have read and agreed with the manuscript before its submission and publication.

Funding

This research received no specific financial support from any funding agency.

Institutional Review Board Statement

Not applicable.

References

[1] Liu, W., Hu, J., & Wang, W. (2020). A Novel Camera Fusion Method Based on Switching Scheme and Occlusion-Aware Object Detection for Real-Time Robotic Grasping. Journal of Intelligent & Robotic Systems, 100(3-4), 791–808. https://doi.org/10.1007/s10846-020-01236-7

[2] Du, G., Wang, K., Lian, S., & Zhao, K. (2021). Vision-based robotic grasping from object localization, object pose estimation to grasp estimation for parallel grippers: a review. Artificial Intelligence Review, 54(3), 1677–1734. https://doi.org/10.1007/s10462-020-09888-5

[3] Watanabe, T. (2022). Robotic Hands for Object Grasping and Manipulation. Journal of the Robotics Society of Japan, 40(5), 363–368. https://doi.org/10.7210/jrsj.40.363

[4] Byambaa, M., Koutaki, G., & Choimaa, L. (2022). 6D Pose Estimation of Transparent Object From Single RGB Image for Robotic Manipulation. IEEE Access, 10, 114897–114906. https://doi.org/10.1109/access.2022.3217811

[5] Park, H. J., An, B. H., Joo, S. B., Kwon, O. W., Kim, M. Y., & Seo, J. (2022). Grasping Time and Pose Selection for Robotic Prosthetic Hand Control Using Deep Learning Based Object Detection. International Journal of Control, Automation and Systems, 20(10), 3410–3417. https://doi.org/10.1007/s12555-021-0449-6

[6] Araki, R., Hirakawa, T., Yamashita, T., & Fujiyoshi, H. (2022). MT-DSSD: multi-task deconvolutional single shot detector for object detection, segmentation, and grasping detection. Advanced Robotics, 36(8), 373–387. https://doi.org/10.1080/01691864.2022.2043183

[7] Muthusamy, R., Huang, X., Zweiri, Y., Seneviratne, L., & Gan, D. (2020). Neuromorphic Event-Based Slip Detection and Suppression in Robotic Grasping and Manipulation. IEEE Access, 8, 153364–153384. https://doi.org/10.1109/access.2020.3017738

[8] Fang, H., Fang, H. S., Xu, S., & Lu, C. (2022). TransCG: A Large-Scale Real-World Dataset for Transparent Object Depth Completion and a Grasping Baseline. IEEE Robotics and Automation Letters, 7(3), 7383–7390. https://doi.org/10.1109/lra.2022.3183256

[9] da Fonseca, V. P., Jiang, X., Petriu, E. M., & de Oliveira, T. E. A. (2022). Tactile object recognition in early phases of grasping using underactuated robotic hands. Intelligent Service Robotics, 15(4), 513–525. https://doi.org/10.1007/s11370-022-00433-7

[10] Wu, Y., Fu, Y., & Wang, S. (2020). Deep instance segmentation and 6D object pose estimation in cluttered scenes for robotic autonomous grasping. Industrial Robot: the international journal of robotics research and application, 47(4), 593–606. https://doi.org/10.1108/ir-12-2019-0259

[11] Li, D., Liu, H., Wei, T., & Zhou, J. (2022). Robotic grasping method of bolster spring based on image-based visual servoing with YOLOv3 object detection algorithm. Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science, 236(3), 1780–1795. https://doi.org/10.1177/09544062211019774

[12] Shaw-Cortez, W., Oetomo, D., Manzie, C., & Choong, P. (2020). Technical Note for "Tactile-Based Blind Grasping: A Discrete-Time Object Manipulation Controller for Robotic Hands". IEEE Robotics and Automation Letters, 5(2), 3475–3476. https://doi.org/10.1109/lra.2020.2977585

[13] Wang, S., Zhou, Z., & Kan, Z. (2022). When Transformer Meets Robotic Grasping: Exploits Context for Efficient Grasp Detection. IEEE Robotics and Automation Letters, 7(3), 8170–8177. https://doi.org/10.1109/lra.2022.3187261

[14] Bai, Q., Li, S., Yang, J., Song, Q., Li, Z., & Zhang, X. (2020). Object Detection Recognition and Robot Grasping Based on Machine Learning: A Survey. IEEE Access, 8, 181855–181879. https://doi.org/10.1109/access.2020.3028740

[15] Fang, W., Chao, F., Lin, C. M., Zhou, D., Yang, L., Chang, X., Shen, Q., & Shang, C. (2021). Visual-Guided Robotic Object Grasping Using Dual Neural Network Controllers. IEEE Transactions on Industrial Informatics, 17(3), 2282–2291. https://doi.org/10.1109/tii.2020.2995142

[16] Yu, Y., Cao, Z., Liu, Z., Geng, W., Yu, J., & Zhang, W. (2020). A two-stream CNN with simultaneous detection and segmentation for robotic grasping. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 52(2), 1167-1181. https://doi.org/10.1109/TSMC.2020.3018757

[17] Sejdiu, V., Pajaziti, A., Rexha, G., Bajrami, X., Rrustemi, E., & Kola, J. (2022). Detection, Recognition, and Grasping of Objects through Artificial Intelligence Using a Robotic Hand. IFAC-PapersOnLine, 55(39), 443–446. https://doi.org/10.1016/j.ifacol.2022.12.077

[18] Biagetti, L., Kochar, A., Cristalli, C., & Boria, S. (2020). Cognitive Grasping System: A Grasping Solution for Industrial Robotic Manipulation using Convolutional Neural Network. Procedia Manufacturing, 51, 32–37. https://doi.org/10.1016/j.promfg.2020.10.006

[19] Valarezo Añazco, E., Rivera Lopez, P., Park, N., Oh, J., Ryu, G., Al-antari, M. A., & Kim, T. S. (2021). Natural object manipulation using anthropomorphic robotic hand through deep reinforcement learning and deep grasping probability network. Applied Intelligence, 51(2), 1041–1055. https://doi.org/10.1007/s10489-020-01870-6

[20] Zhang, H., Yang, D., Wang, H., Zhao, B., Lan, X., Ding, J., & Zheng, N. (2022). REGRAD: A Large-Scale Relational Grasp Dataset for Safe and Object-Specific Robotic Grasping in Clutter. IEEE Robotics and Automation Letters, 7(2), 2929–2936. https://doi.org/10.1109/lra.2022.3142401

[21] Hong, Q. Q., Yang, L., & Zeng, B. (2022). RANET: A Grasp Generative Residual Attention Network for Robotic Grasping Detection. International Journal of Control, Automation and Systems, 20(12), 3996–4004. https://doi.org/10.1007/s12555-021-0929-8

[22] Shashank, T., Hitesh, N., & Gururaja, H. (2022). Application of few-shot object detection in robotic perception. Global Transitions Proceedings, 3(1), 114–118. https://doi.org/10.1016/j.gltp.2022.04.024

[23] Liu, D., Tao, X., Yuan, L., Du, Y., & Cong, M. (2022). Robotic Objects Detection and Grasping in Clutter Based on Cascaded Deep Convolutional Neural Network. IEEE Transactions on Instrumentation and Measurement, 71, 1–10. https://doi.org/10.1109/tim.2021.3129875

[24] Lee, W. C., Zhang, J. Y., & Wei, C. C. (2021). Using an LCD Monitor and a Robotic Arm to Quickly Establish Image Datasets for Object Detection. IEEE Access, 9, 131006-131019. https://doi.org/10.1109/ACCESS.2021.3111314

[25] Chen, P., & Lu, W. (2021). Deep reinforcement learning based moving object grasping. Information Sciences, 565, 62–76. https://doi.org/10.1016/j.ins.2021.01.077

[26] Narang, A., Bajaj, S. B., & Jaglan, V. (2022). Robotic Arm. International Journal of Social Ecology and Sustainable Development, 13(1), 1–13. https://doi.org/10.4018/ijsesd.295967