Beyond RGB-D: Perception of Glass, Mirrors, and See-Through Scenes for Robotics

Salem Ameen* and Hari Sunmukeswar Baskaran

University of Salford, Manchester, United Kingdom

**Corresponding author: Salem Ameen*

Abstract

Perception of transparent and reflective objects remains one of the most challenging problems in robotics because such materials violate assumptions used by conventional vision and depth sensing systems. Reflections, refractions, missing depth, and view-dependent appearance can lead to unreliable scene understanding, affecting navigation, mapping, obstacle avoidance, grasping, and manipulation. Although recent advances in computer vision and deep learning have improved individual perception tasks, many existing approaches are still studied independently, limiting their usefulness in practical robotic systems. This survey reviews recent methods for perception in transparent, reflective, and see-through environments with emphasis on robotics-oriented sensing, geometry recovery, uncertainty estimation, and decision-making. It covers transparent object segmentation, depth completion, layered depth estimation, multimodal sensing, polarization-based geometry recovery, neural rendering, event-based vision, foundation-model-based depth estimation, sim-to-real transfer, and embodied robotic intelligence. Rather than treating these topics as isolated modules, the paper presents a unified perception-pipeline perspective in which sensing, scene understanding, representation transformation, uncertainty reasoning, and task-level interaction are connected. The survey also examines failure cases in SLAM and LiDAR mapping, robotics-centric evaluation metrics, benchmark datasets, computational constraints, and deployment challenges. Finally, it identifies key research gaps and future directions for building robust, efficient, and adaptive robotic perception systems for real-world transparent and reflective environments.

Keywords: transparent object perception; reflective surfaces; robotic vision; RGB-D sensing; depth completion; multimodal fusion; SLAM; uncertainty estimation; neural rendering

1. Introduction

Robotic perception has improved substantially in recent years, allowing autonomous systems to operate in increasingly complex and dynamic environments. However, transparent and reflective objects such as glass, mirrors, glossy surfaces, and transparent containers remain difficult for robotic perception systems. These materials reflect, refract, transmit, and distort light in ways that violate simplified Lambertian assumptions used by many computer vision and depth sensing methods [19], [20]. Consequently, robots may receive incomplete or inaccurate depth measurements, ambiguous boundaries, or false geometric structures. These problems directly affect navigation, obstacle avoidance, mapping, grasping, and scene understanding.

Several research directions have attempted to address these limitations. Learning-based methods such as ClearGrasp, TransCG, and SeeClear improved transparent object perception by using geometry-aware reconstruction, large-scale real-world data, and generative scene transformation [1], [2], [3]. Other work has studied layered depth estimation, fast depth completion, uncertainty-aware depth prediction, and foundation-model-based monocular depth estimation [4], [5], [15], [18], [21]-[23]. In parallel, multimodal approaches using tactile sensing, polarization imaging, neural rendering, and event-based vision provide additional cues that are not available from conventional RGB or RGB-D sensors alone [6]-[8], [24]-[27], [40]-[42].

Despite this progress, many methods are still developed and evaluated as isolated perception components. Real robotic systems, however, require multiple components to operate together in closed-loop pipelines. A depth-completion model may improve reconstruction quality, but the result is useful only if it supports reliable mapping, grasp planning, obstacle avoidance, or task-level reasoning. Transparent and reflective environments also create systematic failures in SLAM, LiDAR mapping, and navigation because reflections and refractions may produce false observations or inconsistent point clouds [9]-[11]. These limitations show that the field requires not only better algorithms but also robotics-centric evaluation frameworks and deployable system designs.

This survey provides a structured review of transparent and reflective object perception for robotics. Its main contribution is to connect methods across segmentation, depth recovery, sensing, reconstruction, uncertainty estimation, benchmarking, and real-time deployment into a unified robotics-oriented perspective. The paper aims to clarify how different techniques contribute to practical robotic tasks, where current limitations remain, and which research directions are most important for robust real-world deployment.

2. Survey Methodology

This survey reviews robotic perception methods designed for transparent, reflective, and see-through environments. The literature search focused on research in computer vision, robotics, depth estimation, SLAM, multimodal sensing, neural rendering, and embodied artificial intelligence. Sources included IEEE Xplore, SpringerLink, ScienceDirect, ACM Digital Library, arXiv, Google Scholar, and proceedings from major venues such as CVPR, ICCV, ECCV, ICRA, IROS, RA-L, TPAMI, WACV, CoRL, and ICML.

The primary coverage period was 2018 to 2026, with selected earlier works included where they provide foundational concepts for polarization imaging, sim-to-real transfer, or robotic manipulation. Search terms included transparent object perception, glass detection, mirror perception, transparent depth estimation, reflection-aware SLAM, transparent object segmentation, RGB-D depth completion, shape from polarization, polarimetric robot perception, event-based vision, neural rendering, reflective reconstruction, uncertainty-aware depth estimation, foundation depth models, sim-to-real transfer, and embodied robotic perception.

Papers were selected based on relevance to robotic perception under non-Lambertian conditions, contribution to geometry recovery or scene understanding, experimental validation, availability of datasets or benchmarks, and practical relevance to navigation, mapping, manipulation, or embodied interaction. Purely computer-graphics papers were excluded unless they contributed directly to reflective reconstruction, neural rendering, physically based simulation, or synthetic data generation for perception. The selected works were grouped into categories covering segmentation, depth completion, layered depth reasoning, foundation models, polarization-based sensing, multimodal fusion, neural rendering, event-based vision, SLAM failure analysis, benchmarks, metrics, real-time deployment, and sim-to-real transfer.

3. Taxonomy of Perception Approaches

3.1. Unified Perception Pipeline

Non-Lambertian perception should not be viewed as a single-step prediction problem. Recent work suggests that reliable robotic perception usually requires a sequence of interconnected stages involving sensing, region localization, representation transformation, geometry recovery, uncertainty reasoning, and task-level decision-making. ClearGrasp, for example, showed that estimating intermediate signals such as masks, surface normals, and boundary cues can be more effective than directly predicting depth for transparent objects [1]. TransCG

demonstrated how large-scale real-world RGB-D data and learned refinement can convert corrupted sensor observations into depth representations useful for manipulation [2]. SeeClear further extended this direction by transforming transparent regions into geometry-consistent opaque representations that can be processed by existing depth models [3].

Figure 1 illustrates the unified perception-pipeline view adopted in this survey. The pipeline shifts attention from isolated modules to their interactions. This is important because real-world failures often occur not because a single module is weak, but because the overall perception-to-action pipeline is poorly integrated. A robotics-oriented review must therefore consider how segmentation, depth estimation, uncertainty, mapping, and control influence each other.

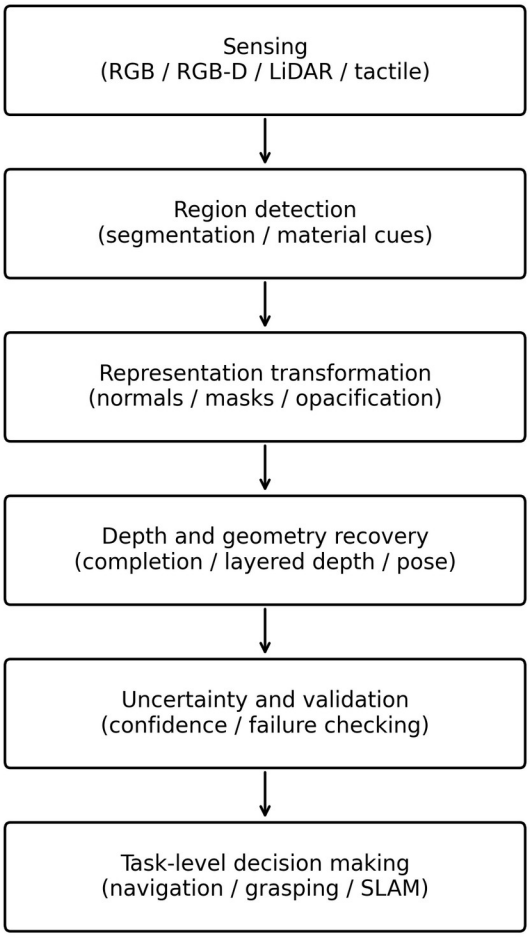

Figure 1. Unified perception pipeline for transparent and reflective object understanding in robotics.

3.2. Transparent Object Segmentation and Scene Understanding

Transparent object segmentation is challenging because transparent surfaces often borrow texture, colour, and illumination from the background. Boundaries may be weak, and visible appearance may vary with viewpoint and lighting. Early learning-based segmentation methods therefore focused on boundary-aware semantic segmentation and transparent-scene datasets. Trans10K and TransLab provided important benchmarks for

transparent object segmentation in the wild, enabling models to learn transparent boundaries under diverse lighting and occlusion conditions [31].

Reflective and mirror-rich environments introduce related but distinct challenges. MirrorNet and Mirror3D addressed mirror recognition, mirror-aware depth refinement, and geometric consistency in reflective scenes [28], [29]. These works show that segmentation is not merely a semantic labelling problem. In robotics, segmentation acts as an intermediate step that supports obstacle detection, mapping, depth correction, grasp planning, and safe interaction with the environment.

3.3. Depth Completion and Layered Depth Estimation

Depth estimation for transparent and reflective objects is difficult because RGB-D sensors may return missing, distorted, or background-dominated measurements. ClearGrasp addressed this problem by combining segmentation, surface normals, and boundary cues to improve transparent object depth estimation for manipulation [1]. TransCG expanded the field by introducing a large-scale real-world dataset and a depth-completion baseline for transparent object grasping [2]. FDCT later emphasized the importance of high-speed transparent depth completion for real-time robotic perception [18].

In see-through scenes, a single depth value per pixel may be insufficient because multiple surfaces can exist along the same viewing ray. Layered depth estimation therefore attempts to recover both transparent foreground surfaces and background structures [4]. DepthFocus extends this idea by treating depth estimation as a controllable process, allowing models to focus on different layers depending on task requirements [5]. For robotics, this is important because navigation may prioritize the nearest transparent obstacle, while manipulation may require understanding objects or surfaces behind transparent materials.

3.4. Foundation Models and Generalized Depth Estimation

Foundation-model-based depth estimation has improved generalization across diverse environments. Models such as Depth Anything V2, ZoeDepth, and Marigold use large-scale pretraining, metric-depth transfer, transformer architectures, and diffusion-based learning to produce robust monocular depth estimates [21]-[23]. These models are attractive for robotics because they can operate from RGB images and often generalize better than task-specific supervised models.

However, foundation depth models are not automatically reliable in transparent and reflective environments. Their predictions may appear visually plausible while failing to represent true geometry near glass, mirrors, or specular objects. ConFiDeNet addresses part of this limitation by jointly estimating depth and uncertainty, allowing systems to identify low-confidence regions [15]. Fast-FoundationStereo highlights another important direction: adapting foundation-level stereo perception to real-time constraints [16]. In robotics, the next step is to connect foundation-model predictions with uncertainty calibration, material awareness, and task-level safety.

3.5. Polarization-Based Geometry Recovery

Polarization imaging provides material and surface-orientation information that conventional RGB cameras cannot capture. Foundational work showed that surface orientation can be recovered from diffuse polarization patterns [37]. Later transparent surface modelling approaches used pairs or multiple views of polarization images to reduce ambiguity in surface normal estimation [38].

Deep shape-from-polarization methods combine physical reflection models with neural networks to estimate surface normals more accurately under complex lighting [30]. Polarized 3D also showed that polarization cues can refine coarse depth estimates and improve geometric detail [39]. Recent surveys on polarimetric imaging argue that polarization sensing is particularly valuable for robotics because it captures reflection, refraction, and

material cues that are unavailable in RGB-only perception [8]. The main limitation is that polarization cameras and calibration pipelines add hardware cost and deployment complexity.

3.6. Multimodal Fusion and Active Tactile Perception

Transparent object perception is often unreliable when using a single sensing modality. RGB images may provide ambiguous appearance, RGB-D sensors may produce missing depth, and LiDAR may introduce false reflections. Multimodal fusion addresses this limitation by combining complementary signals. Visual-tactile fusion methods have improved transparent object grasping by integrating camera observations with tactile feedback, enabling robots to infer object boundaries and contact geometry when visual cues are weak [6].

Active tactile reconstruction extends this idea by allowing the robot to interact with the object and progressively refine its understanding. ACTOR, for example, uses tactile interaction for category-level transparent object reconstruction [7]. These approaches are especially relevant for manipulation, where physical contact can provide reliable geometric information even when vision is uncertain. Multimodal fusion also includes combinations of RGB, depth, polarization, thermal, LiDAR, and event-camera data, but the key research challenge remains how to fuse heterogeneous signals without increasing latency or system complexity beyond practical deployment limits.

3.7. Neural Rendering and Reflection-Aware Reconstruction

Neural rendering has become increasingly relevant for transparent and reflective scene reconstruction. Neural Radiance Fields introduced continuous volumetric scene representations for novel-view synthesis [43], while broader neural rendering research has developed methods for learning appearance, geometry, and view-dependent effects from images [24]. Ref-NeRF improves the modelling of reflective and specular surfaces through structured view-dependent appearance [25]. Neuralangelo enhances high-fidelity neural surface reconstruction from RGB observations [26].

Reflection-aware methods such as NeRF-Casting further attempt to model consistent reflections in neural scene representations [27]. Physically based differentiable rendering frameworks such as Mitsuba 2 are also important because they enable more realistic simulation of light transport, reflection, and material interactions [44]. Neural reflectance fields support appearance acquisition and material modelling [45]. Although these methods are often computationally expensive, they are becoming increasingly useful for robotic mapping, digital twins, synthetic data generation, and evaluation of perception systems in transparent and reflective environments.

3.8. Event-Based Vision for Dynamic Robotic Perception

Event-based cameras record asynchronous brightness changes rather than full frames, providing high temporal resolution, low latency, and high dynamic range [40]. These properties are useful for robotic systems that operate under rapid motion, glare, or extreme lighting, all of which are common near reflective and transparent surfaces. Event-based high-dynamic-range reconstruction has shown that event cameras can recover visual information in conditions where conventional frame-based cameras may saturate or blur [41].

Ultimate SLAM combined events, images, and inertial measurements for robust visual SLAM under high-speed and HDR scenarios [42]. While event cameras do not solve transparent-object perception alone, they provide complementary temporal information that can improve robustness in dynamic scenes. Future robotic systems may combine event sensing with depth prediction, material-aware segmentation, and uncertainty estimation to handle fast motion and reflection-related failures more reliably.

3.9. Robotics Manipulation and Embodied Intelligence

Perception for transparent and reflective objects becomes most valuable when it improves robotic action. Visual-tactile fusion and active tactile reconstruction connect perception with manipulation by using interaction to

resolve visual ambiguity [6], [7]. Earlier large-scale learning for hand-eye coordination demonstrated the value of data-driven visual control for robotic grasping [32]. More recent embodied multimodal models such as PaLM-E integrate vision, language, and robotic reasoning, suggesting a future direction where perception, task understanding, and action planning are jointly optimized [33].

For transparent and reflective environments, embodied intelligence is especially important because passive observation is often insufficient. A robot may need to move, touch, change viewpoint, use active illumination, or query multiple sensors to reduce uncertainty. This shifts the research problem from static scene analysis toward active perception and closed-loop interaction.

4. Failure Modes, Benchmarking, and Evaluation

4.1. Failure-Case Analysis in SLAM and Mapping

Transparent and reflective surfaces can cause systematic failures in localization and mapping. Visual SLAM systems may interpret mirror reflections as real objects, creating inconsistent trajectories and duplicate geometry [9]. LiDAR-based mapping can also be affected when glass or mirrors create false returns, transmitted measurements, or reflected points [10]. Recent plane-optimization and plane-SLAM approaches show that reflected points, real objects, and transmitted signals can coexist in the same scene, making robust interpretation difficult [11].

These failures matter because robotic safety depends on reliable environment understanding. A visually accurate depth map is not enough if it creates unsafe navigation or incorrect grasp planning. Surveying transparent and reflective perception therefore requires explicit analysis of failure cases, not only average performance metrics.

4.2. Robotics-Centric Evaluation Frameworks

Standard computer vision metrics do not always capture practical robotic usefulness. ClearPose provides a large-scale transparent object benchmark for depth completion, segmentation, and object pose estimation [12]. Booster focuses on depth estimation for specular and transparent surfaces and highlights performance degradation in material-specific regions [13]. TRICKY 2025 extends this direction by targeting failure-prone scenes with glass, mirrors, occlusions, and material-aware evaluation [14].

Robotics-centric evaluation should therefore consider not only pixel-level accuracy but also downstream outcomes such as grasp success, collision avoidance, mapping consistency, inference speed, and uncertainty calibration. A perception system that is slightly less accurate on a benchmark may still be preferable if it is faster, better calibrated, or safer under uncertainty.

4.3. Comparative Analysis of Methods

Table 1 summarizes representative methods across transparent and reflective object perception. The comparison highlights differences in sensing modality, core task, deployment feasibility, and robotics relevance. No single method dominates across all conditions; method selection depends on the target task, sensor constraints, and real-time requirements.

Table 1. Comparative analysis of representative transparent and reflective object perception methods.

Method	Year	Sensor requirement	Core task	Strengths	Limitations	Deployment	Robotics relevance	Ref.
ClearGrasp	2020	RGB-D camera	Transparent depth completion	Strong geometry recovery for transparent objects	Depends on RGB-D correction and object assumptions	Near real-time	Robotic grasping	[1]
TransCG	2022	RGB-D camera	Transparent depth completion	Large real-world dataset and robust grasping baseline	Focused mainly on indoor RGB-D setups	Real-time	Industrial manipulation	[2]
SeeClear	2026	RGB camera	Transparent monocular depth	Plug-and-play transformation for foundation depth models	Diffusion pipeline may be computationally expensive	Offline / near real-time depending on implementation	Transparent scene perception	[3]
Layered depth	2025	RGB / RGB-D	Multi-layer depth estimation	Models foreground transparent surfaces and background structure	Layer selection remains task-dependent	Offline	See-through scene understanding	[4]
DepthFocus	2025	RGB camera	Controllable depth estimation	Selects task-relevant depth layers	Requires control objective and careful validation	Offline	Task-aware depth reasoning	[5]
Visual-tactile fusion	2023	Vision + tactile sensors	Transparent object grasping	Improves geometry through contact feedback	Requires tactile hardware and interaction	Near real-time	Robotic manipulation	[6]
ACTOR	2023	Visual + tactile sensors	Active transparent reconstruction	Improves shape estimation through interaction	Slow and interaction-dependent	Offline / interactive	Manipulation and exploration	[7]
FDCT	2023	RGB-D camera	Fast depth completion	High-speed transparent object depth completion	May reduce fine geometric detail	Real-time	Embedded perception	[18]
ConFiDeNet	2026	RGB camera	Depth + uncertainty estimation	Provides confidence for risk-aware reasoning	Higher computational cost and calibration demands	Near real-time	Safety-aware perception	[15]
Depth Anything V2	2024	RGB camera	Foundation monocular depth	Strong zero-shot generalization	Weak explicit material reasoning	Near real-time	General robotic perception	[21]
ZoeDepth	2023	RGB camera	Metric monocular depth	Strong metric-depth transfer	Limited handling of reflections and transparency	Near real-time	Navigation and mapping	[22]
Marigold	2024	RGB camera	Diffusion-based depth	Fine-grained depth reconstruction	High latency	Offline	Scene understanding	[23]
Deep shape from polarization	2022	Polarization camera	Shape from polarization	Strong surface-normal estimation	Requires specialized hardware and calibration	Offline	Surface geometry reconstruction	[30]
Mirror3D	2021	RGB-D camera	Mirror-aware depth	Reflective depth understanding	Mirror-specific assumptions	Near real-time	Indoor robotics	[29]
Ultimate SLAM	2018	Event camera + RGB + IMU	Robust visual SLAM	HDR and high-speed robustness	Complex sensor synchronization	Near real-time	Autonomous navigation	[42]
Ref-NeRF	2022	Multi-view RGB	Reflective neural rendering	High-quality reflective appearance modelling	Computationally expensive	Offline	Digital twins and simulation	[25]
Domain randomization	2017	Simulation environment	Sim-to-real transfer	Improves generalization from simulation to reality	Requires extensive randomized training	Near real-time after training	Robot learning	[34]
Dynamics randomization	2018	Robot simulator	Robotic control transfer	Improves adaptation to real robots	Depends on simulator fidelity and tuning	Near real-time after training	Robotic control	[35]

4.4. Datasets and Benchmarks

The quality and diversity of datasets strongly influence progress in transparent and reflective object perception. Synthetic data can provide complete ground truth but may suffer from domain gaps. Real-world datasets capture deployment conditions but are expensive to annotate, especially when depth sensors fail on transparent surfaces. Table 2 summarizes major datasets and benchmarks used in the field.

Table 2. Datasets and benchmarks for transparent and reflective object perception.

Dataset / benchmark	Year	Primary task	Data type	Modalities	Scale / scope	Key features	Limitations	Robotics applicability	Ref.
ClearGrasp dataset	2020	Transparent depth completion	Synthetic + real	RGB-D	50k+ synthetic images and real test images	Transparent masks, normals, geometry	Limited real-world training diversity	Robotic grasping	[1]
TransCG	2022	Transparent depth completion	Real-world	RGB-D	57,715 RGB-D images across 130 scenes	Large-scale real-world transparent-object data	Indoor RGB-D focus	Industrial grasping	[2]
Trans10K / TransLab	2020	Transparent segmentation	Real-world	RGB	10k annotated images	Boundary-aware transparent segmentation	No depth or pose labels	Scene understanding and navigation	[31]
ClearPose	2022	Transparent pose estimation	Real-world	RGB-D	Pose benchmark for transparent objects	Accurate pose annotations	Primarily pose-focused	Robotic manipulation	[12]
Booster	2024	Specular / transparent depth benchmark	Benchmark dataset	RGB-D	Material-aware depth evaluation	Highlights hard transparent and reflective regions	Limited deployment diversity	Benchmarking and evaluation	[13]
TRICKY 2025	2025	Transparent and reflective evaluation	Benchmark / challenge	RGB-D and robotics sensors	Multi-scenario evaluation	Failure-prone materials and occlusions	Emerging benchmark	Robotics benchmarking	[14]
SeeClear-396k	2026	Transparent monocular depth	Synthetic	RGB	396k transparent-opaque paired renderings	Aligned depth, normals, masks and opacity supervision	Synthetic domain gap	Transparent scene perception	[3]
Mirror3D	2021	Mirror-aware depth	Real-world	RGB-D	Mirror-aware depth annotations	Reflective scene understanding	Mirror-specific	Indoor robotics and navigation	[29]
Polarized 3D data	2015	Polarization depth enhancement	Controlled experimental	Polarization + depth	Multi-view polarization captures	Supports polarization-guided reconstruction	Requires specialized setup	3D reconstruction	[39]
ACTOR tactile data	2023	Active tactile reconstruction	Real-world	Visual + tactile	Interactive tactile exploration	Combines touch and visual perception	Expensive data collection	Robotic manipulation	[7]
Ultimate SLAM HDR data	2018	Event-based SLAM	Real-world	Event camera + RGB + IMU	HDR and high-speed motion sequences	Supports robust SLAM under extreme conditions	Complex multimodal synchronization	Autonomous navigation	[42]

4.5. Evaluation Metrics

Transparent and reflective object perception requires multiple complementary metrics. Segmentation metrics evaluate object localization, depth metrics measure reconstruction quality, SLAM metrics measure mapping and trajectory consistency, and task metrics evaluate practical robotic outcomes. Table 3 summarizes common metrics and their robotics relevance.

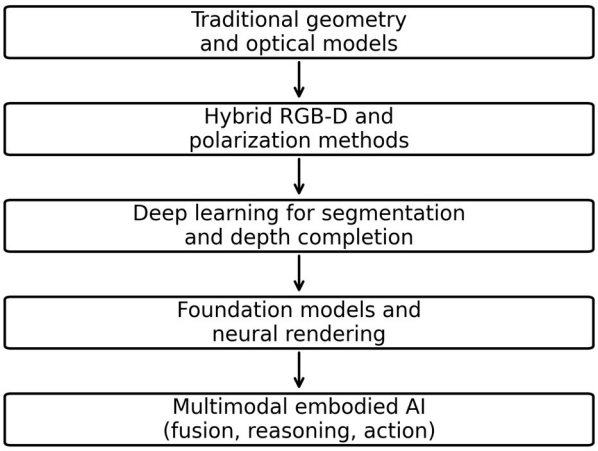

Figure 2. Evolution of transparent and reflective object perception methods from geometric reasoning toward multimodal embodied AI.

Table 3. Common evaluation metrics for transparent and reflective object perception.

Metric	Application area	Purpose	Advantages	Limitations	Robotics relevance
Intersection over Union (IoU)	Segmentation	Measures overlap between predicted and ground-truth regions	Widely used and easy to interpret	Sensitive to boundary errors	Obstacle and scene understanding
Boundary F-score	Transparent segmentation	Measures boundary precision and recall	Useful for thin transparent edges	Sensitive to annotation quality	Transparent boundary detection
Pixel accuracy	Segmentation	Measures percentage of correctly classified pixels	Simple and fast	Can be misleading for imbalanced data	Real-time perception evaluation
RMSE	Depth estimation	Measures depth prediction error magnitude	Captures large errors	Sensitive to outliers	Navigation and depth reliability
AbsRel	Depth estimation	Measures relative depth error	Common monocular depth metric	Affected by scale inconsistency	Robotic scene understanding
Threshold accuracy (delta)	Depth estimation	Measures predictions within a threshold	Useful for robust depth evaluation	Does not fully capture geometry consistency	Reliable depth perception
Chamfer distance	3D reconstruction	Measures similarity between point clouds	Useful for geometry evaluation	Sensitive to sparse or noisy points	3D mapping and reconstruction
ADD / ADD-S	6DoF pose estimation	Measures pose estimation accuracy	Widely used in manipulation benchmarks	Affected by object symmetry	Grasping and manipulation
ATE / RPE	SLAM	Measures trajectory and localization consistency	Effective for SLAM benchmarking	Environment-dependent	Autonomous navigation
Grasp success rate	Robotic manipulation	Measures successful grasp attempts	Directly task relevant	Hardware/setup dependent	Transparent object grasping
Collision rate	Navigation	Measures collision events	Safety-relevant	Scenario-dependent	Safe navigation
Inference speed (FPS)	Deployment	Measures processing rate	Important for embedded robotics	High speed does not ensure accuracy	Real-time robotic systems
Uncertainty calibration error	Uncertainty estimation	Measures reliability of confidence predictions	Supports risk-aware decisions	Difficult to standardize	Safety-critical robotics

Evaluation should not rely on a single metric. A robotics-oriented system may require acceptable segmentation accuracy, reliable depth, calibrated uncertainty, stable SLAM, sufficient inference speed, and successful task execution. Future benchmarks should therefore evaluate perception reliability and task-level performance together rather than treating them as separate problems.

5. Deployment Constraints and Sim-to-Real Transfer

5.1. Real-Time and Embedded Deployment Constraints

Deploying perception models on robots introduces constraints that are often absent in offline benchmarks. Robots operate under strict latency, memory, power, and hardware constraints. High-performing models based on large transformers, diffusion pipelines, or neural rendering can provide strong accuracy but may be unsuitable for closed-loop navigation or manipulation if inference is slow [16], [21], [23], [25], [26].

Size, weight, and power constraints are especially important for drones, mobile robots, and embedded platforms. Lightweight transparent-obstacle detection using sensor fusion has demonstrated that combining Time-of-Flight and ultrasonic sensors with compact neural models can support real-time obstacle awareness onboard small aerial robots [17]. Figure 3 presents a generalized embedded deployment pipeline. Efficient models such as FDCT also show that fast depth completion can support high-frame-rate perception for transparent objects [18].

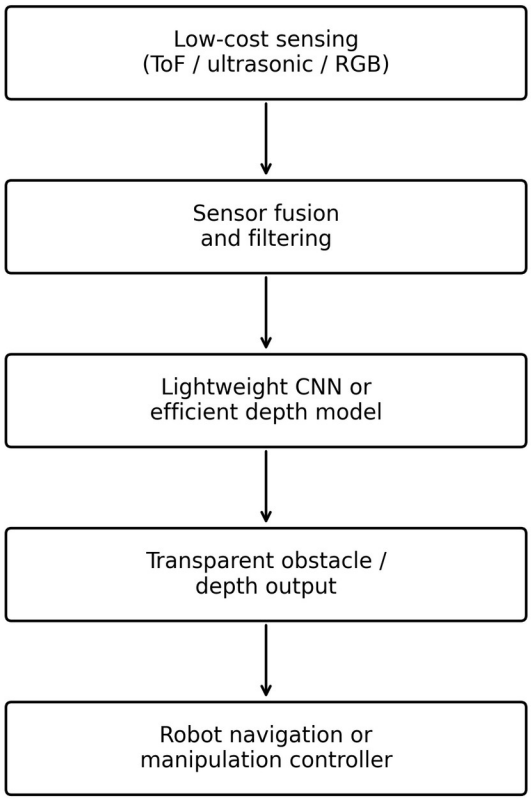

Figure 3. Generalized real-time deployment pipeline for transparent obstacle perception using lightweight sensing and efficient neural inference.

The key deployment challenge is balancing accuracy, speed, uncertainty, and robustness. A high-accuracy offline model may fail in dynamic environments if predictions arrive too late. Conversely, a fast model may be unsafe if it is poorly calibrated near transparent obstacles. Practical systems therefore require end-to-end optimization, including efficient feature extraction, lightweight fusion, robust confidence estimation, and minimal post-processing latency.

5.2. Sim-to-Real Transfer and Synthetic Data Generation

Collecting and annotating real transparent-object datasets is expensive because sensors often fail to capture reliable geometry. Synthetic rendering can provide complete annotations for depth, normals, masks,

transparency, and object pose. Domain randomization showed that models trained in simulation can generalize to the real world when sufficient variability is introduced during training [34]. Dynamics randomization extended this idea to robotic control transfer by varying physical parameters during simulation [35].

For transparent and reflective perception, synthetic data is particularly useful because physically based rendering can generate controlled material interactions that are difficult to annotate in real scenes [24], [44]. ClearGrasp, SeeClear, and related methods use synthetic rendering or paired transformations to provide geometry-consistent supervision [1], [3]. However, sim-to-real transfer remains challenging because real transparent and reflective scenes contain complex illumination, sensor noise, occlusion, and material variation. Future datasets should combine synthetic scale with real-world validation, multimodal sensing, temporal sequences, and task-level robotic interaction.

6. Research Gaps and Future Directions

First, the field lacks unified robotic perception systems that jointly integrate segmentation, depth completion, layered depth reasoning, SLAM, uncertainty estimation, and task-level decision-making. Existing work often improves one module, but practical robots need coordinated pipelines. Future research should evaluate whether improved perception actually improves navigation safety, grasp success, and mapping reliability.

Second, transparent and reflective scene understanding remains weak under changing illumination, dynamic backgrounds, strong occlusion, and variable material properties. Robust models should combine visual appearance, geometry, polarization, tactile feedback, temporal cues, and uncertainty rather than relying on RGB or RGB-D alone. Active perception, including viewpoint selection and tactile exploration, is likely to become increasingly important.

Third, uncertainty estimation is underused in robotic decision-making. Many systems provide deterministic outputs even when predictions near glass or mirrors are unreliable. Future work should calibrate uncertainty, distinguish aleatoric from epistemic uncertainty, and connect confidence maps to safe navigation, cautious manipulation, and recovery behaviours [15].

Fourth, real-time deployment remains unresolved for many high-performing approaches. Foundation models, diffusion-based depth estimation, and neural rendering can improve reconstruction quality, but their computational demands limit use on embedded platforms [21], [23], [25], [26]. Lightweight multimodal fusion, model compression, distillation, and hardware-aware design are therefore essential for deployable systems.

Fifth, datasets and benchmarks require broader robotics-oriented coverage. Existing datasets are valuable but often focus on isolated tasks, static scenes, or limited sensor configurations. Future benchmarks should include transparent and reflective materials under dynamic lighting, diverse viewpoints, temporal sequences, multimodal sensors, manipulation interactions, and safety-critical navigation scenarios.

7. Conclusion

This survey reviewed recent perception methods for transparent, reflective, and see-through environments with a focus on robotics. It covered transparent object segmentation, depth completion, layered depth estimation, foundation models, multimodal sensing, polarization-based geometry recovery, neural rendering, event-based vision, SLAM failures, benchmarking, uncertainty estimation, real-time deployment, sim-to-real transfer, and embodied robotic intelligence.

The review shows that transparent and reflective perception cannot be solved through isolated improvements in individual modules alone. Robust robotic operation requires integrated perception pipelines that connect sensing, representation transformation, depth and geometry recovery, uncertainty reasoning, and task-level

action. Current systems still face major challenges in generalization, computational efficiency, calibrated uncertainty, dataset diversity, and reliable deployment in dynamic real-world environments.

Future progress is likely to come from lightweight multimodal perception systems, uncertainty-aware decision-making, robotics-centric benchmarks, synthetic-to-real training pipelines, and embodied intelligence systems that combine perception with active interaction. Building such systems is essential for enabling robots to navigate, manipulate, and reason safely in environments containing glass, mirrors, transparent objects, and other non-Lambertian materials.

References

[1] S. S. Sajjan, M. Moore, M. Pan, G. Nagaraja, J. Lee, A. Zeng, and S. Song, "ClearGrasp: 3D shape estimation of transparent objects for manipulation," in Proc. IEEE International Conference on Robotics and Automation (ICRA), pp. 3634-3642, 2020, doi: 10.1109/ICRA40945.2020.9197518.

[2] H. Fang, H.-S. Fang, S. Xu, and C. Lu, "TransCG: A large-scale real-world dataset for transparent object depth completion and a grasping baseline," IEEE Robotics and Automation Letters, vol. 7, no. 3, pp. 7383-7390, 2022, doi: 10.1109/LRA.2022.3183102.

[3] X. Wang, Y. He, J. Shi, J. Lu, Y. Yang, Y. Jiang, and C. Jiang, "SeeClear: Reliable transparent object depth estimation via generative opacification," arXiv:2603.19547, 2026. [Preprint].

[4] H. Wen, Y. Zuo, V. Subramanian, P. Chen, and J. Deng, "Seeing and seeing through the glass: Real and synthetic data for multi-layer depth estimation," in Proc. IEEE/CVF International Conference on Computer Vision (ICCV), pp. 6715-6725, 2025.

[5] J. Min, J. Kim, C.-H. Min, M. Kim, Y. Jeon, and M. Choi, "DepthFocus: Controllable depth estimation for see-through scenes," arXiv:2511.16993, 2025. [Preprint].

[6] S. Li, H. Yu, W. Ding, H. Liu, L. Ye, C. Xia, X. Wang, and X.-P. Zhang, "Visual-tactile fusion for transparent object grasping in complex backgrounds," IEEE Transactions on Robotics, vol. 39, no. 5, pp. 3838-3856, 2023, doi: 10.1109/TRO.2023.3286071.

[7] P. K. Murali, B. Porr, and M. Kaboli, "Touch if it is transparent! ACTOR: Active tactile-based category-level transparent object reconstruction," in Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 10792-10799, 2023, doi: 10.1109/IROS55552.2023.10341680.

[8] C. Taglione, C. Mateo, and C. Stolz, "Polarimetric imaging for robot perception: A review," Sensors, vol. 24, Art. no. 4440, 2024, doi: 10.3390/s24144440.

[9] P. Herbert, J. Wu, Z. Ji, and Y.-K. Lai, "Benchmarking visual SLAM methods in mirror environments," Computational Visual Media, vol. 10, no. 2, pp. 215-241, 2024.

[10] X. Zhao, Z. Yang, and S. Schwertfeger, "Mapping with reflection: Detection and utilization of reflection in 3D LiDAR scans," in Proc. IEEE International Conference on Robotics and Automation (ICRA), 2020.

[11] Y. Li, X. Zhao, and S. Schwertfeger, "Detection and utilization of reflections in LiDAR scans through plane optimization and plane SLAM," Sensors, vol. 24, no. 15, Art. no. 4794, 2024, doi: 10.3390/s24154794.

[12] X. Chen, H. Zhang, Z. Yu, A. Opipari, and O. C. Jenkins, "ClearPose: Large-scale transparent object dataset and benchmark," in Proc. European Conference on Computer Vision (ECCV), pp. 381-396, 2022.

[13] P. Z. Ramirez, A. Costanzino, F. Tosi, M. Poggi, S. Salti, S. Mattoccia, and L. Di Stefano, "Booster: A benchmark for depth from images of specular and transparent surfaces," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 46, no. 1, pp. 85-102, 2024, doi: 10.1109/TPAMI.2023.3323858.

[14] P. Z. Ramirez et al., "TRICKY 2025 challenge on monocular depth from images of specular and transparent surfaces," in Proc. IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), pp. 3311-3322, 2025.

[15] O. Susladkar, R. Pawar, C. Sehgal, S. Ujjawal, and S. Mittal, "Confidence through parallel attention for depth and uncertainty estimation in dynamic environments," in Proc. IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2026.

[16] B. Wen, S. Dewan, and S. Birchfield, "Fast-FoundationStereo: Real-time zero-shot stereo matching," arXiv:2512.11130, 2025. [Preprint].

[17] M. Hopkins, V. Murali, V. Kumar, and C. J. Taylor, "Real-time glass detection and reprojection using sensor fusion onboard aerial robots," IEEE Robotics and Automation Letters, 2023.

[18] T. Li, Z. Chen, H. Liu, and C. Wang, "FDCT: Fast depth completion for transparent objects," IEEE Robotics and Automation Letters, vol. 8, no. 9, pp. 5823-5830, 2023, doi: 10.1109/LRA.2023.3300544.

[19] J. Jiang, G. Cao, J. Deng, T.-T. Do, and S. Luo, "Robotic perception of transparent objects: A review," IEEE Transactions on Artificial Intelligence, vol. 5, no. 6, pp. 2547-2567, 2024, doi: 10.1109/TAI.2023.3326120.

[20] F. Li, J. Ma, H.-N. Liang, Z. Tian, Z. Wu, T. Wen, and D. Liu, "A comprehensive survey of specularity detection: State-of-the-art techniques and breakthroughs," Artificial Intelligence Review, vol. 58, Art. no. 218, 2025, doi: 10.1007/s10462-025-11233-7.

[21] L. Yang, B. Kang, Z. Huang, X. Xu, J. Feng, and H. Zhao, "Depth Anything V2," arXiv:2406.09414, 2024. [Preprint].

[22] S. Bhat, R. Alhashim, and P. Wonka, "ZoeDepth: Zero-shot transfer by combining relative and metric depth," in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 18156-18165, 2023.

[23] K. Keetha, A. Karazija, A. Sanakoyeu, D. P. Kingma, P. Vedaldi, and C. Rupprecht, "Marigold: Repurposing diffusion-based image generators for monocular depth estimation," in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024.

[24] S. Tewari, M. Zollhoefer, C. Kim, P. Garrido, F. Bernard, P. Perez, and C. Theobalt, "State of the art on neural rendering," Computer Graphics Forum, vol. 39, no. 2, pp. 701-727, 2020, doi: 10.1111/cgf.14022.

[25] B. Verbin, P. Hedman, B. Mildenhall, F. Lafortune, D. Novotny, and P. Srinivasan, "Ref-NeRF: Structured view-dependent appearance for neural radiance fields," in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5481-5490, 2022.

[26] T. Li, M. Goel, Q. Chen, and C. Theobalt, "Neuralangelo: High-fidelity neural surface reconstruction," in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 8456-8465, 2023.

[27] Y. Chen, H. Xu, M. Wang, X. Yang, and G. Lin, "NeRF-Casting: Improved view-dependent appearance with consistent reflections," arXiv:2405.14871, 2024. [Preprint].

[28] X. Yang, H. Qi, and J. Jia, "MirrorNet: Bio-inspired framework for mirror recognition," IEEE Transactions on Image Processing, vol. 30, pp. 8600-8615, 2021, doi: 10.1109/TIP.2021.3116090.

[29] X. Luo, Y. Xiong, K. Lin, and G. Wang, "Mirror3D: Depth refinement for mirror surfaces," in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 15177-15186, 2021.

[30] C. Ba, T. Xiang, S. Shao, C. Hane, and J. Yu, "Deep shape from polarization," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 10, pp. 6936-6950, 2022, doi: 10.1109/TPAMI.2021.3110906.

[31] X. Yang, H. Qi, C. Wang, and J. Jia, "Segmenting transparent objects in the wild," in Proc. European Conference on Computer Vision (ECCV), pp. 696-711, 2020.

[32] S. Levine, P. Pastor, A. Krizhevsky, and D. Quillen, "Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection," International Journal of Robotics Research, vol. 37, no. 4-5, pp. 421-436, 2018, doi: 10.1177/0278364917710318.

[33] D. Driess, F. Xia, M. Sajjadi, et al., "PaLM-E: An embodied multimodal language model," in Proc. International Conference on Machine Learning (ICML), 2023.

[34] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel, "Domain randomization for transferring deep neural networks from simulation to the real world," in Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 23-30, 2017.

[35] X. B. Peng, M. Andrychowicz, W. Zaremba, and P. Abbeel, "Sim-to-real transfer of robotic control with dynamics randomization," in Proc. IEEE International Conference on Robotics and Automation (ICRA), pp. 3803-3810, 2018.

[36] A. Atapour-Abarghouei and T. P. Breckon, "Real-time monocular depth estimation using synthetic data with domain adaptation via image style transfer," in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2800-2810, 2018.

[37] S. K. Nayar, X.-S. Fang, and T. Boult, "Recovery of surface orientation from diffuse polarization," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 19, no. 7, pp. 690-705, 1997, doi: 10.1109/34.598228.

[38] M. Miyazaki, K. Hara, and K. Ikeuchi, "Transparent surface modeling from a pair of polarization images," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 28, no. 1, pp. 73-87, 2006, doi: 10.1109/TPAMI.2006.18.

[39] Y. Kadambi, V. Taamazyan, B. Shi, and R. Raskar, "Polarized 3D: High-quality depth sensing with polarization cues," in Proc. IEEE International Conference on Computer Vision (ICCV), pp. 3370-3378, 2015.

[40] G. Gallego, T. Delbruck, G. Orchard, et al., "Event-based vision: A survey," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 1, pp. 154-180, 2022, doi: 10.1109/TPAMI.2020.3008413.

[41] H. Rebecq, R. Ranftl, V. Koltun, and D. Scaramuzza, "High speed and high dynamic range video with an event camera," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 43, no. 6, pp. 1964-1980, 2021, doi: 10.1109/TPAMI.2019.2963386.

[42] A. Rosinol Vidal, H. Rebecq, T. Horstschaefer, and D. Scaramuzza, "Ultimate SLAM? Combining events, images, and IMU for robust visual SLAM in HDR and high-speed scenarios," IEEE Robotics and Automation Letters, vol. 3, no. 2, pp. 994-1001, 2018, doi: 10.1109/LRA.2018.2793357.

[43] B. Mildenhall, P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "NeRF: Representing scenes as neural radiance fields for view synthesis," in Proc. European Conference on Computer Vision (ECCV), pp. 405-421, 2020.

[44] W. Jakob, D. Vicini, T. Zeltner, et al., "Mitsuba 2: A retargetable forward and inverse renderer," ACM Transactions on Graphics, vol. 38, no. 6, Art. no. 203, 2019, doi: 10.1145/3355089.3356498.

[45] A. Bi, J. Xu, K. Sunkavalli, et al., "Neural reflectance fields for appearance acquisition," arXiv:2008.03824, 2020. [Preprint].