IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

TORM: Transparent Objects Reconstruction and Manipulation with Multi-View Segmentation

Qiyuan Qiao[1], Fuling Lin[1], Huibin Zhao[1], Bowen Xu[1], Zhiqiang Chen[1], Dong Xu[2], Peng Lu[1]

***Abstract*—Transparent objects are common in daily life and industry, necessitating that robots be able to perceive and manipulate them. The physical properties of reflection and refraction pose challenges for accurately reconstructing the 3D geometry of transparent objects. Conventional methods, which rely on simultaneous estimation of background ambient light and complex refraction fields, lack robustness in real-world scenes, thereby impeding robotic grasping performance. To address this issue, this paper proposes TORM, a novel framework for robust reconstruction and manipulation of multiple transparent objects. TORM focuses on semantic information from transparent objects and employs multi-view segmentation masks to constrain a self-supervised multi-object deep marching tetrahedra (DMTet-Multi) 3D fitting process. To mitigate the risk of the geometry representation getting stuck in suboptimal solutions during multi-transparent-object reconstruction, we design a novel loss function that prevents marching tetrahedra from crossing boundaries. By applying a connectivity determination strategy to the fitted mesh, transparent objects can be processed in parallel by a grasp perception network, predicting the end-effector configuration for grasp tasks. Real-world experiments demonstrate that TORM achieves an 88.8% grasping success rate in multi-transparent-object grasping tasks.**

***Index Terms*—Perception for grasping and manipulation; Computer vision for automation; Deep learning for visual perception**

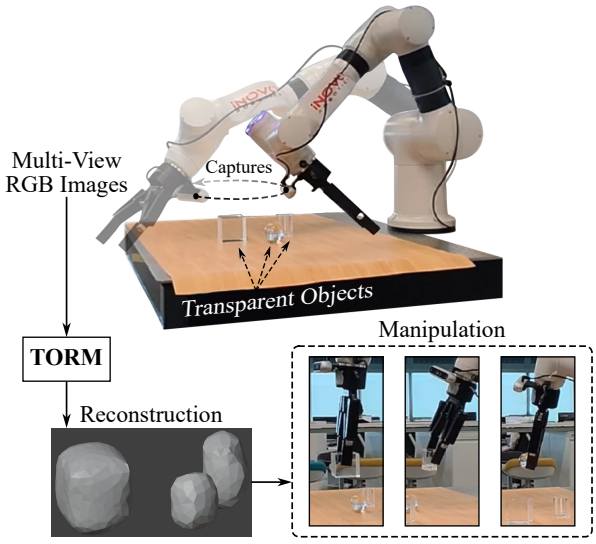

Fig. 1. **TORM** accepts multi-view RGB inputs and utilizes differentiable mesh representations to achieve robust 3D reconstruction and pose estimation for grasp tasks.

I. Introduction

THE reconstruction and grasping of transparent objects are critical for robotics [1], augmented reality [2], and laboratory automation [3], [4], where accurate 3D surface estimation underpins effective manipulation. Given 2D images from a camera, robots must reconstruct the 3D geometry of transparent objects and execute interactive grasping tasks. However, the lack of distinct features, coupled with surface reflections in scenes with multiple transparent objects, complicates shape perception and requires semantic understanding for object differentiation and grasp pose planning.

Prior approaches to address these perception challenges include multi-view stereo [5], [6] and neural radiance fields [1], [7], which rely on visual cues but struggle with computationally expensive refraction modeling. Specialized depth sensors, such as time-of-flight [8] or learning-based depth completion [9]–[12], offer alternatives but introduce additional costs and often lack sufficient annotated training data.

The complexity of modeling light propagation, including parameters like surface glossiness and refractive index, further hinders geometric estimation, as these are computationally intensive and sensitive to disturbances such as vibrations or temperature fluctuations [8]. Critically, existing methods often fail to distinguish individual transparent objects from each other or the background, providing only whole-scene geometric data that necessitates re-perception for precise grasp planning [10].

To address these fundamental challenges, our approach is motivated by two key observations. Recent advances in large-scale semantic segmentation models enable robust extraction of transparent object silhouettes, providing stable geometric cues that remain consistent across varying optical conditions. Moreover, existing approaches face critical limitations when handling multiple transparent objects: sequential reconstruction of individual objects increases computational cost, while reconstructing the entire scene at once often leads to sub-optimal solutions that fail to capture all objects and do not separate individual instances. Based on these observations,

Manuscript received: June 24, 2025; Revised September 21, 2025; Accepted October 23, 2025. This paper was recommended for publication by Associate Editor T. Welschehold and Editor Markus Vincze upon evaluation of the Associate Editor and Reviewers' comments. This work was supported by General Research Fund under grant no. 17204222. *(Qiyuan Qiao and Fuling Lin contributed equally to this work.) (Corresponding author: Peng Lu)*

[1]Qiyuan Qiao, Fuling Lin, Huibin Zhao, Bowen Xu, Zhiqiang Chen and Peng Lu are with Department of Mechanical Engineering, The University of Hong Kong, Hong Kong SAR, China. lupeng@hku.hk

[2]Dong Xu is with School of Computing and Data Science, The University of Hong Kong, Hong Kong SAR, China.

Digital Object Identifier (DOI): see top of this page.

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

we propose **TORM**, a novel pipeline for robust Transparent Object Reconstruction and Manipulation. Illustrated in Fig. 1, a differentiable hybrid-encoded mesh representation is optimized using the supervision of object silhouettes and subsequently used for grasp pose generation.

Based on segmented silhouettes from different views, we propose multi-object deep marching tetrahedra (DMTet-Multi), which can constrain a deep 3D conditional implicit surface reconstruction network for diverse objects. Our DMTet-Multi encodes a discretized signed distance function (SDF) and optimizes a deformable tetrahedral grid to generate the explicit surface mesh representation. By applying multi-view silhouette constraints to the deformable mesh and SDF field, TORM emphasizes mask edges and distributes contributions across viewpoints, preventing low-quality views from compromising reconstruction fidelity.

Though the deformable mesh representation can theoretically handle topological changes, it struggles when fitting multiple disconnected objects due to complex loss landscapes that trap the optimization in suboptimal local minima. To address this, we propose a progressive loss to shrink the mesh, initialized outside the reconstruction volume, inward to envelop all target objects. This approach ensures that each object's topology is first captured as disconnected structures before surface refinement, thereby avoiding solutions that only cover a subset of objects.

Transparent object geometry is extracted in a single reconstruction pass by applying a connectivity-based separation of the fitted mesh, yielding distinct 3D point cloud representations for each object. These segmented point clouds are then processed, along with their semantic identifiers and local background context, by a grasp perception network to operate in parallel to predict end-effector configurations for all targets simultaneously. This strategy eliminates the need for repeated per-object re-perception, enabling stable and efficient planning of grasp poses for each transparent item within a scene.

Our main contributions can be summarized as follows:

1) A novel framework, TORM, is proposed for robust reconstruction and manipulation of multiple transparent objects, including the DMTet-Multi model for simultaneous multi-object reconstruction and a connectivity determination strategy to separate the fitted mesh.
2) A specialized loss with a progressively decreasing envelope constraint is proposed to prevent DMTet-Multi from becoming trapped in suboptimal solutions during multi-transparent-object reconstruction.
3) Extensive experiments and comparative evaluations against state-of-the-art methods are conducted in both simulation and real-world settings, validated by a newly constructed multi-view RGB dataset of transparent object scenes spanning both environments.

II. Related Work

A. Transparent Object Segmentation

The silhouettes, masks, or occluding contours are widely used in 3D reconstruction and surface estimation [13], [14]. Early works [15], [16] utilize silhouettes for reconstructing the shapes of reflective or refractive objects, which necessitate accurate silhouette segmentation. Conventional silhouette detection methods, such as background subtraction [17] or edge detection with the Canny operator [18], are often disrupted by the background transmitted pattern. Recent approaches [19], [20] advance the detection and segmentation of non-Lambertian objects with deep learning methods. Conventional methods are often trained on limited, domain-specific datasets [4] and struggle with significant domain shifts in diverse real-world environments, frequently requiring extensive retraining on new data. Foundation models, such as the segment anything model (SAM) [21], demonstrate remarkable improvements in silhouette detection, primarily due to their large-scale pretraining on diverse datasets, strong zero-shot generalization capabilities, and unified visual representations that capture both global context and fine-grained details without requiring domain-specific optimization for transparent objects. In the field of 2D-image-to-3D-model estimation, the surface normal is also an important feature besides the mask. Some research [10], [12] predicts the normal information for the 3D shape supervision. However, the ground truth (GT) of the surface normal of transparent objects usually relies on computer graphics simulation, causing an unignorable domain gap in physical world manipulation tasks. In contrast, our approach relies solely on silhouette features for 3D reconstruction by leveraging the power of a foundation model, thereby circumventing the domain gap issues inherent in normal estimation while still providing sufficient geometric constraints to recover the 3D shape of transparent objects.

B. Deformable Tetrahedral Mesh Reconstruction

Selecting an appropriate representation for transparent object reconstruction presents unique challenges that demand specific capabilities from the underlying geometric model. Mesh-based representation methods [22], [23] provide the explicit manipulable geometry information for direct robotic tasks. Implicit neural representations [24]–[26] offer greater topological flexibility for 3D reconstruction. Implicit approaches such as DeepSDF [24] excel at representing arbitrary topologies and enable end-to-end differentiable optimization. Recent hybrid approaches combine the advantages of both explicit and implicit representations [27], [28]. The deep marching tetrahedra (DMTet) approach [29] provides an elegant solution by embedding an implicit SDF within a tetrahedral grid and employing differentiable marching tetrahedra [30] to extract explicit meshes with coarse point cloud inputs. It supports arbitrary topological structures, which are essential for modeling complex transparent objects. Besides, this representation enables differentiable loss propagation from silhouette supervision to the geometric model, making it suitable for multi-view constraint-based 3D reconstruction. In this work, we utilize DMTet to effectively reconstruct multiple objects from multi-view silhouettes, while preserving the geometric details necessary for robotic grasping.

C. Transparent Object Grasping Techniques

Robotic grasping of transparent objects presents unique challenges due to their optical properties interfering with

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

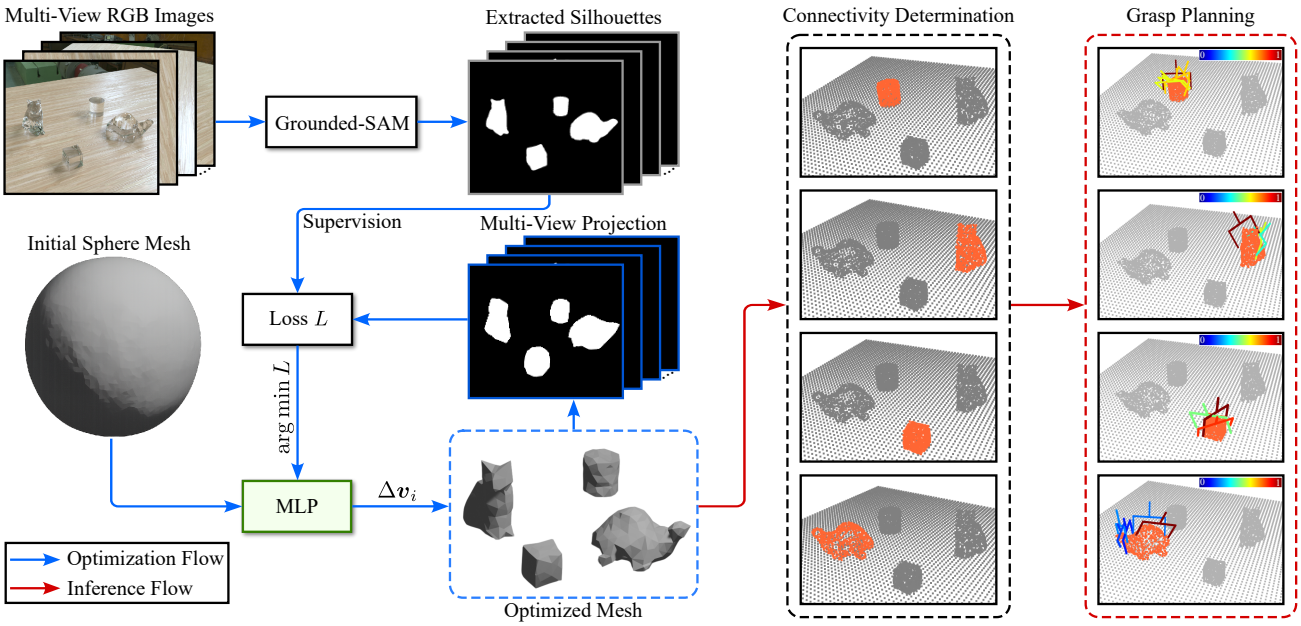

Fig. 2. **Overview of the TORM pipeline**, consisting of three main components: (1) multi-view semantic segmentation, (2) DMTet-Multi for geometric reconstruction, and (3) object-level grasp planning based on split point cloud features via a connectivity determination strategy. The multi-view semantic segmentation provides the silhouette information for the DMTet-Multi reconstruction, which is then used to generate point clouds for grasp planning.

conventional perception systems. Neural radiance fields and 3D gaussian splatting approaches [1], [31] often generate artifacts or incorrectly fit transparent object shapes to background textures, while depth completion methods [10], [12] struggle with unrealistic edge interpolation and unobserved viewpoint gaps. In the grasping domain, research typically employs point cloud feature extraction with PointNet++ [32] backbone networks as seen in GraspNet [33], but the pipeline of sampling depth maps into point clouds introduces additional complexities for transparent objects due to sensor limitations. In contrast, our approach bridges the gap between accurate 3D reconstruction and effective manipulation by sampling high-quality point clouds directly from our tetrahedral mesh reconstructions. Unlike methods that rely on single-view depth maps, which inherently lose information and impede globally optimal grasp pose estimation, our approach reconstructs complete, topologically accurate 3D models based on only RGB information. This allows us to overcome challenges, including inconsistent depth readings and missing geometric features. The integration of our deformable tetrahedral representation with point cloud-based grasping techniques creates a robust end-to-end solution for transparent object manipulation.

III. Method

The schematic in Fig. 2 presents the TORM workflow, which unfolds in three key stages: (1) extracting multi-view semantic silhouettes by harnessing cues specific to transparent objects (Sec. III-A); (2) reconstructing the geometry of multiple objects via DMTet-Multi, guided by a tailored loss function (Sec. III-B); (3) generating object-level grasp poses based on the reconstructed 3D models (Sec. III-C).

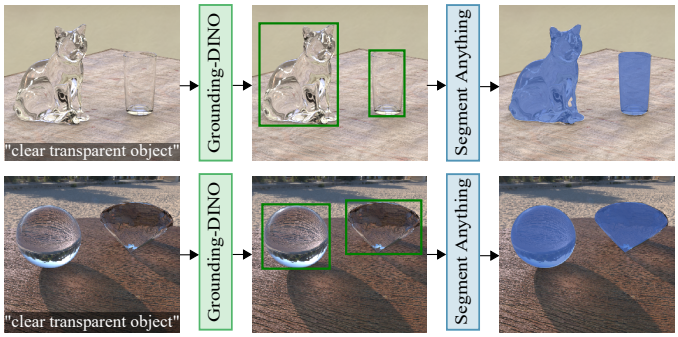

Fig. 3. **Silhouettes extraction of transparent objects.** Grounding DINO predicts bounding boxes according to a text prompt from RGB images, and the boxes serve as the prompts for silhouette extraction in SAM.

A. Transparent Object Segmentation

Fresnel reflectance, a physical phenomenon observable at transparent object boundaries, can be exploited to robustly extract silhouettes with semantic labels. Transparent objects manipulate incoming light through refraction and reflection, producing a highly reliable signal at the object boundary: a bright, view-dependent rim caused by a sharp increase in surface reflectance as the incident angle approaches grazing incidence. This provides reliable supervisory signals for 3D reconstruction and shows strong robustness to noise and camera calibration errors, which motivates us to develop a dedicated segmentation approach for transparent objects.

Building on this physical insight, the silhouette extraction employs a two-stage strategy as shown in Fig. 3. While task-specific models for transparent object segmentation remain viable, the impressive generalization and zero-shot transfer abilities of recent foundation models enable more efficient solutions. In the first stage, an open-set object detector, Ground-

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

ing DINO [34], detects arbitrary objects based on text prompts. By leveraging vision-language pretraining, prompts such as "clear glass" and "transparent objects" are utilized to scan multi-view images, enabling the integration of visual cues with semantic understanding for effective localization in complex scenarios. These bounding boxes serve as spatial prompts for the subsequent stage, in which SAM [21] generates accurate silhouettes based on the detected regions.

The extracted silhouettes serve dual purposes: providing supervisory signals for optimizing topological structures of transparent objects while filtering out background clutter to facilitate downstream tasks such as grasp planning. Anchoring perception to physically-grounded edge features bypasses the complexities of full light-path modeling while preserving the information essential for 3D shape recovery.

B. 3D Multi-Object Reconstruction

Our proposed DMTet-Multi extends the conventional DMTet framework to address multi-object scenarios, overcoming the premature convergence to object subsets caused by local optima in deformable meshes. DMTet-Multi introduces a strategic decoupling of topology formation from geometric refinement through a temporally adaptive envelope constraint, facilitating robust reconstruction of multiple disconnected objects. The framework utilizes a tetrahedral grid where each vertex $\boldsymbol{v}_i$ is characterized by an SDF value $s_i \in \mathbb{R}$ and a deformation vector $\Delta\boldsymbol{v}_i \in \mathbb{R}^3$, with a three-layer perceptron $f(\boldsymbol{v}_i)$ employed to predict these parameters under silhouette-based supervision, thereby enabling concurrent optimization of surface geometry and topological structure.

Through this differentiable architecture, binary object silhouettes are rendered and gradients are propagated backward to optimize the geometry encoded in per-vertex SDF values and deformation parameters. The initial mask loss function is formulated as:

$$L_{\text{mask}}^{\text{init}} = \frac{1}{N}\sum_{i=1}^{N}\left(\hat{M}_i - M_i\right)^2 \tag{1}$$

where $\hat{M}_i$ and M_i denote the predicted and GT binary silhouettes at pixel position i, respectively, and N represents the total number of pixels. This pixel-wise loss computation encompasses the entire image domain, with resulting gradients backpropagated to update both SDF values and deformation parameters.

However, (1) presents a critical limitation: mesh extension toward disconnected objects necessitates traversal through background regions, consequently increasing the loss value and constraining the optimization within local minima. To circumvent this topological barrier, the proposed method initializes an oversized tetrahedral sphere and incorporates a temporally adaptive envelope constraint coefficient α_{gt}, reformulating the envelope mask loss as:

$$L_{\text{mask}} = \frac{1}{N}\sum_{i=1}^{N}\left(\hat{M}_i - \alpha_{\text{gt}} M_i\right)^2 \tag{2}$$

where α_{gt} constitutes a scalar coefficient that undergoes temporal decay according to:

$$\alpha_{\text{gt}}(t) = \begin{cases} \alpha_{\max} - \frac{\alpha_{\max}-\alpha_{\min}}{T}\cdot t & \text{if } 0 \le t < T \\ \alpha_{\min} & \text{if } t \ge T \end{cases} \tag{3}$$

where t denotes the current iteration, $\alpha_{\max}$ and $\alpha_{\min}$ define the constraint bounds, and T specifies the decay duration. This progressive relaxation strategy enables the deformable mesh to initially establish correct topological connectivity before converging to precise geometric configurations, thereby maintaining individual object supervision throughout the optimization process.

The eikonal loss is utilized to constrain the implicit encoded SDF in DMTet-Multi, enhancing the surface quality and regularization of the SDF field. With vertex SDF value s_i and displaced vertex position $\boldsymbol{v}_i'$, the surface is detected based on the sign transitions of s_i. For an edge with endpoints $\boldsymbol{v}_i$ and $\boldsymbol{v}_j$, when $\text{sign}(s_i) \neq \text{sign}(s_j)$, the zero-crossing position is computed via linear interpolation: $\boldsymbol{v}_0 = \frac{\boldsymbol{v}_i' s_j - \boldsymbol{v}_j' s_i}{s_j - s_i}$. For each interpolated point $\boldsymbol{v}_0$ in the set of zero SDF values V_0, we apply eikonal regularization [35]:

$$L_{\text{eikonal}} = \frac{1}{|V_0|}\left(\|\nabla_{\boldsymbol{v}} f_s(\boldsymbol{v}_0)\| - 1\right)^2 \tag{4}$$

where $\nabla_{\boldsymbol{v}} f_s$ denotes the gradient of the SDF output from the network f with respect to the vertex position $\boldsymbol{v}$.

The reconstructed mesh encompassing multiple objects is optimized using the combination of the mask loss L_{mask} and the eikonal regularization L_{eikonal}:

$$L = \sum^{N_{\text{view}}} \lambda_{\text{mask}} L_{\text{mask}} + \lambda_{\text{eikonal}} L_{\text{eikonal}} \tag{5}$$

where λ_{mask} and λ_{eikonal} are hyperparameters. N_{view} is the number of supervision images involved in an iteration.

C. Object-Level Grasp Planning from Reconstructed Meshes

A mesh-based grasp planning method is proposed that generates object-specific grasp poses for multiple transparent objects simultaneously, eliminating the requirement for additional depth sensors while enabling targeted manipulation. In contrast, previous approaches like GraspNet [33] and AnyGrasp [36] require external depth cameras and can only generate scene-level grasp configurations.

This object-level planning approach leverages the multi-object meshes from DMTet-Multi reconstruction, which provides complete geometric information beyond single-viewpoint depth maps. A graph-based topological decomposition partitions the unified mesh into object-level components: each triangular face forms a node in an adjacency graph, with connected-component analysis identifying topologically disjoint regions corresponding to individual objects.

The segmented meshes are uniformly sampled into point clouds with equal point count for each object, enabling parallel batch processing. A planar support surface representation is added for collision-aware planning. By adapting GraspNet to operate on these object-level point clouds, TORM can achieve the simultaneous grasp generation for all objects through single-pass inference.

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

TABLE I
PERFORMANCE OF DIFFERENT LOSSES ON THE SYNTHETIC DATASET

Mask Loss in (5)	Silhouettes		Average	
	GT	Seg.	Iteration	IoU
$L_{\text{mask}}^{\text{init}}$	✓		—	0.151
L_{mask}	✓		8000	0.876
L_{mask}		✓	8000	0.808

IV. EXPERIMENTS

Evaluations on our self-constructed synthetic and real-world robotic grasping scenarios validate the effectiveness of the proposed TORM compared to existing methods.

A. Experiment Setup

Due to the lack of suitable datasets for multiple transparent object reconstruction and manipulation, we create a new benchmark consisting of 12 synthetic scenarios (Syn1-12) and 8 real-world scenarios (RW1-8). Scenarios Syn1-9 and RW1-6 feature moderately spaced objects, whereas Syn10-12 and RW7-8 present densely packed transparent objects to better reflect real-world applications such as laboratory automation and household transparent object handling tasks. In these crowded scenarios with frequent object occlusions across multiple viewpoints, the inter-object spacing is reduced to about half the object width or less, while still allowing minimal clearance for gripper access.

1) Synthetic Dataset: To validate the reconstruction component of our pipeline, a synthetic dataset is generated in Blender, comprising 12 scenarios representative of challenging tabletop grasping of transparent objects. Each scenario contains 2-10 randomly placed objects with highly specular and transparent material properties, forming complex arrangements that feature challenges like irregular geometry, textureless areas, strong specularity, and heavy occlusion. For each of the 12 scenarios, we render 100 multi-view images. To emulate the operational viewpoints of a camera in a typical robotic grasping setup, the camera poses are sampled on an upper hemisphere with a radius of 0.5m. The images are rendered at a resolution of 2160×1440 pixels using the Cycles engine with 4096 samples per pixel, ensuring high-fidelity, noise-free results. The virtual camera's field of view is set to 91.5° to align with the RGB stream of the Intel RealSense D455 depth camera used in our real-world experiments. Sample images for each simulated scenario are shown in Fig. 4, where the occlusion ratio (OCC), defined as the proportion of images across multiple viewpoints in which at least one object is partially occluded by another, is also indicated for each scenario. GT depth maps and object masks are extracted directly from the rendering pipeline.

2) Real-World Dataset: For the real-world dataset, each scenario is captured with 50 RGB images at a resolution of 1280×720 from varying viewpoints using the RealSense D455. After hand-eye calibration, the camera's coordinates in the world coordinate system can be obtained by the robot arm configuration through kinematic calculations. The experiment on physical grasping reported in Sec. IV-E shows the list of real-world datasets.

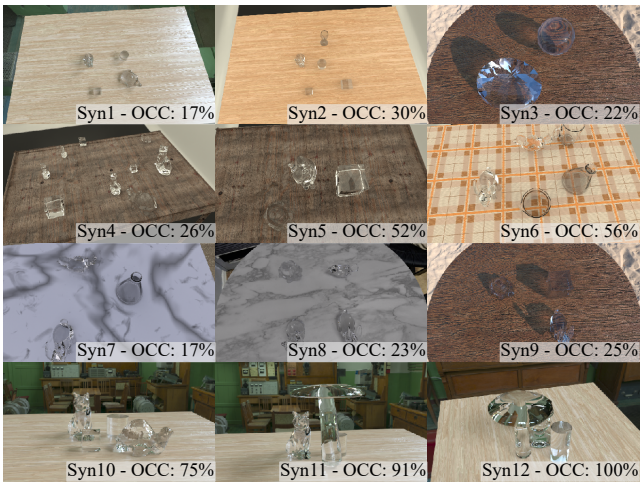

Fig. 4. **Synthetic multiple transparent objects dataset with occlusion**. These scenarios are generated using Blender rendering and feature various transparent objects, rich background textures, and crowded scenarios with frequent object occlusions across multiple viewpoints.

B. Evaluation on Envelop Constraint

Table I presents a quantitative comparison of different envelope constraint strategies for multi-object transparent object reconstruction on the synthetic dataset. The intersection over union (IoU) is used as the evaluation metric to measure the overlap between the 2D projection of the fitted mesh and the supervision masks (GT or segmented silhouettes). The reported IoU values in the table are computed by averaging the projection IoU results from the final five training cycles across our experiments. The initial mask loss, $L_{\text{mask}}^{\text{init}}$, with GT supervision achieves a low average IoU of 0.151, indicating poor performance due to its tendency to converge to suboptimal solutions. On our synthetic dataset, $L_{\text{mask}}^{\text{init}}$ with GT masks frequently suffers from excessive mesh contraction, often causing the mesh to collapse entirely by about 10k iterations, leading to a failure of the fitting process. In such cases, the final IoU for the affected experiment is recorded as 0. To mitigate this collapse, a lower learning rate is applied for $L_{\text{mask}}^{\text{init}}$ in our experiments. In contrast, our proposed envelope constraint, L_{mask}, significantly improves performance, reaching an IoU of 0.876 with GT supervision and 0.808 with segmented silhouette supervision after 8k iterations. Fig. 5 further demonstrates the training dynamics on the Syn5, showing IoU and loss curves for different strategies. The performance of L_{mask} with segmented masks is nearly as effective as GT supervision, underscoring the strength of our DMTet-Multi in achieving high-quality reconstruction.

Fig. 6 visually compares the reconstruction process in scenarios with multiple transparent objects. The top and bottom rows illustrate deformable mesh fitting for two-object and three-object scenarios, respectively, where gray patterns represent target silhouettes and pale violet patterns show 2D mesh projections. With $L_{\text{mask}}^{\text{init}}$, the mesh often shrinks excessively and converges to a local optimum, reconstructing only a subset of objects. For instance, in a two-object scenario, it successfully fits only one object, and after 2k iterations, the deformation field struggles to expand to the second object

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

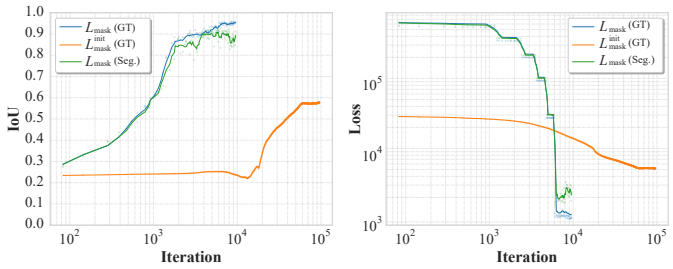

Fig. 5. **Training dynamics comparison in IoU and loss on Syn5.** The labels indicate the type of envelope loss and the corresponding supervision source in the DMTet-Multi training.

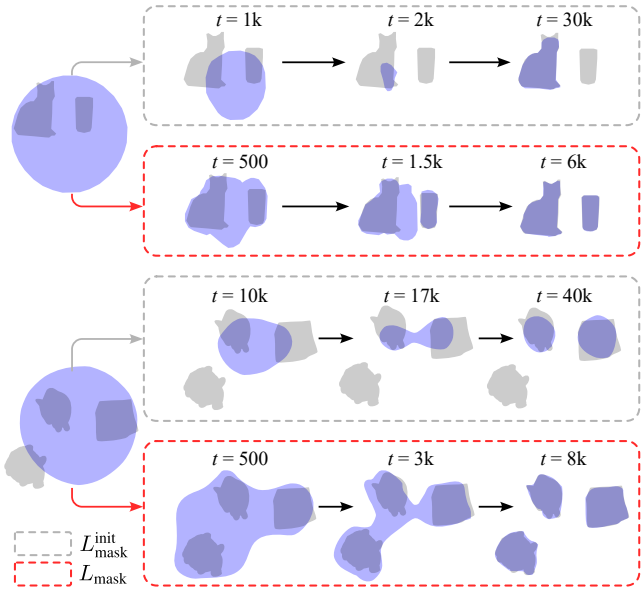

Fig. 6. **Effects of different mask losses.** The top and bottom blocks show deformable mesh fitting for two-object and three-object scenarios, respectively. t represents the iteration steps. Gray patterns are silhouette targets, and pale violet patterns are 2D mesh projections. $L^{\text{init}}_{\text{mask}}$ causes excessive shrinking and local optima, while the proposed L_{mask} ensures correct topology generation.

TABLE II
DEPTH ESTIMATION RESULTS ON THE TORM SYNTHETIC DATASET

Method	RMSE(↓)	AbsRel(↓)	MAE(↓)	$\delta_{1.05}$(↑)	$\delta_{1.10}$(↑)	$\delta_{1.25}$(↑)
Dex-NeRF [1]	0.065	0.090	0.036	66.8	76.1	89.3
3DGS [37]	0.036	0.053	0.021	71.2	87.4	**97.1**
MODEST [38]	0.306	0.739	0.297	0.8	1.5	3.5
MVTrans (5-view) [5]	0.222	0.516	0.199	9.3	17.6	37.2
TORM (Ours)	**0.030**	**0.031**	**0.013**	**88.7**	**93.2**	97.0

due to a steep loss gradient caused by non-target regions. Our envelope constraint, L_{mask}, addresses this by gradually reducing the constraint coefficient α_{gt}, enabling simultaneous reconstruction of all objects with correct topology.

C. Reconstruction Comparisons on Synthetic Dataset

We evaluate the depth estimation performance of our TORM against Dex-NeRF [1], 3DGS [37], MODEST [38], and MVTrans [5] on the synthetic dataset. Depth maps are rendered from reconstructed 3D models and assessed using GT masks to focus on transparent object regions. The evaluation metrics include root mean square error (RMSE), absolute relative error

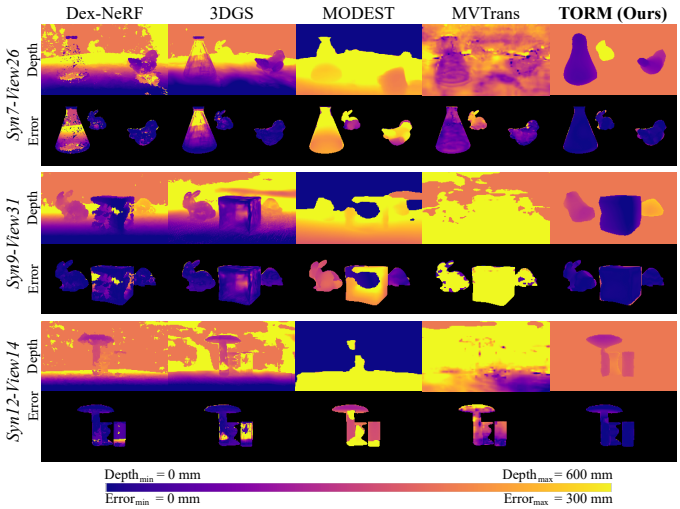

Fig. 7. **Depth estimation results from different methods.** Depth maps and pixel-wise absolute error maps are calculated, and four views from the experiment are shown.

(AbsRel), mean absolute error (MAE), and accuracy thresholds $\delta_{1.05}$, $\delta_{1.10}$, and $\delta_{1.25}$, which represent the percentage of pixels with estimated depth within 105%, 110%, and 125% of the GT, respectively. Table II summarizes the results averaged across all synthetic scenarios, each with 100 viewpoints. TORM outperforms baselines, achieving the lowest errors (RMSE: 0.030, AbsRel: 0.031, MAE: 0.013) and the highest $\delta_{1.05}$ accuracy (88.7%). As shown in Fig. 7, TORM produces artifact-free reconstructions without requiring prior depth information, even on smooth and highly transparent surfaces.

D. Qualitative Analysis of Grasp Planning

Fig. 8 shows the grasp planning results for 2 simulation scenario and 4 real-world scenarios. After connectivity detection is completed, each individual object is uniformly sampled with a sufficient number of densely spaced sampling points. The cyan grasp pose notations correspond one-to-one with the reconstructed objects, and their distribution avoids collision with the table or other objects in the scene while ensuring the grasp quality. Even in the relatively cluttered scenarios (Syn12, RW7, and RW8), TORM can reconstruct objects accurately and generate end-effector grasp poses that are physically feasible and executable, highlighting the robustness of our framework under challenging conditions.

The average time consumption of TORM is analyzed in the real-world test. The semantic silhouette extraction takes 29.0 s to process 50 RGB frames. The deformable mesh optimization achieves a sufficiently good geometric representation in 5k steps, taking 177.3 s. The module for grasp planning can complete the generation of grasp poses for every target transparent object in the scenarios in 14.5 s.

E. Real-World Test for Manipulation

To evaluate the practical effectiveness of our approach in robotic manipulation tasks, we conduct grasping experiments using hardware consisting of an Inovo robotic arm, a Robotiq

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

TABLE III
GRASPING SUCCESS RATES FOR TRANSPARENT OBJECTS IN REAL-WORLD SCENARIOS

Scenario	OCC	Objects	Dex-NeRF [1]	GraspNet [33]	TORM (Ours)	Scenario	OCC	Objects	Dex-NeRF	GraspNet	TORM (Ours)
RW1	0%	Cylinder	*failed*	3/10	**10/10**	RW5	28%	Ball Cup (Opaque)	*failed*	0/10 8/10	**10/10**
RW2	0%	Cup (Opaque)	*failed*	8/10	**10/10**	RW6	2%	Ball Heart Ice Cube Toothpick Box	*failed*	*failed*	**6/10**
RW3	10%	Cylinder Hexagonal Prism	*failed*	3/10 2/10	**9/10**	RW7	46%	Ball Cylinder Toothpick Box	*failed*	*failed*	**9/10**
RW4	16%	Ball Cube Toothpick Box	*failed*	*failed*	**9/10**	RW8	42%	Cylinder Pyramidal Frustum Toothpick Box	*failed*	*failed*	**8/10**

In real-world scenarios, OCC denotes the average occlusion ratio over 10 trials.

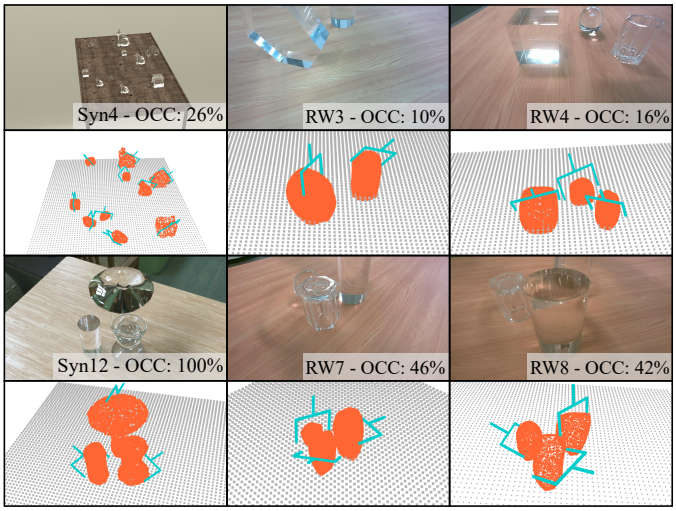

Fig. 8. **Grasp planning examples of TORM in challenging multi-object occlusion scenes.** For each separated object in the reconstructed scenes, a grasp pose with the highest score is retained.

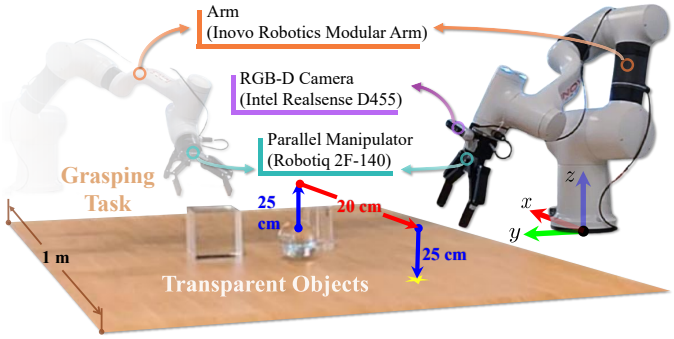

Fig. 9. **Settings of our physical world validation system.** The camera, gripper, and reconstruction space are all calibrated and aligned to the world coordinate system with the robot base as the origin.

gripper, and a RealSense D455 camera. Fig. 9 shows the third-person view of our experimental setup. Supplementary video material is provided to illustrate the experimental setup and the grasping tasks. All 8 real-world scenarios in our self-constructed dataset are used for evaluation. For each scenario, 10 trials are conducted to evaluate the grasping success rate. A successful grasp is defined as lifting the target object 25 cm above the table surface, followed by a 20 cm lateral translation, before lowering and releasing it back onto the table.

For the baseline methods that lack the capability to simultaneously generate grasp poses for multiple objects in a scenario, we evaluate their grasp success rates for individual objects. The real-world manipulation test results are provided in Table III. The proposed TORM framework achieves an average 88.8% success rate across all test cases. Notably, for scenarios with occlusion ratios exceeding 40% (RW7 and RW8), our method still attains high success rates. Dex-NeRF [1] struggles with severe artifacts in reconstruction due to noise in RGB images and camera pose errors during real-world robotic experiments. GraspNet [33], which incorporates an external depth camera, incurs additional sensor costs and calibration challenges while struggling with highly transparent objects with low texture.

Despite the overall high performance of TORM, a few failure cases are observed. These failures primarily stem from three scenarios: (1) occasional gripper slippage on the smooth glass surfaces (e.g., RW3 and RW4); (2) grasping very small objects, such as the 'Heart' in RW6, where the gripper prematurely contacts the table, inadvertently displacing the target object; (3) conservative collision checking in densely cluttered scenes (e.g., RW7 and RW8), which can incorrectly prune viable candidate poses, leaving some objects without grasping configurations. These occasional failures highlight specific challenges but do not undermine the overall effectiveness of TORM, providing valuable insights for future improvements.

V. CONCLUSION

In this paper, we presented TORM, a novel framework that addresses transparent object reconstruction and manipulation challenges by leveraging multi-view semantic segmentations to guide a self-supervised DMTet-Multi fitting process. Our approach focuses on semantic information and silhouette features to achieve robust reconstruction. TORM's key contributions include a novel loss function that prevents marching tetrahedra boundary crossings and a connectivity determination

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

strategy that enables parallel grasp planning for multiple transparent objects. Comprehensive experiments demonstrated TORM's effectiveness with an 88.8% grasping success rate, representing a significant advancement in robotic transparent object manipulation capabilities. We believe that our work contributes to advancing robotic capabilities in understanding and manipulating transparent objects in real-world grasping scenarios. Building on this foundation, future work will aim to improve the runtime efficiency of TORM, enabling faster reconstruction and grasp planning to better support real-time robotic applications and extend to dynamic environments.

REFERENCES

[1] J. Ichnowski, Y. Avigal, J. Kerr, and K. Goldberg, "Dex-nerf: Using a neural radiance field to grasp transparent objects," in *Conference on Robot Learning (CoRL)*, 2021, pp. 526–536.

[2] A. Torres-Gómez and W. Mayol-Cuevas, "Recognition and reconstruction of transparent objects for augmented reality," in *IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, 2014, pp. 129–134.

[3] L. C. O. Tiong, H. J. Yoo, N. Kim, C. Kim, K.-Y. Lee, S. S. Han, and D. Kim, "Machine vision-based detections of transparent chemical vessels toward the safe automation of material synthesis," *npj Computational Materials*, vol. 10, no. 1, p. 42, 2024.

[4] S. Eppel, H. Xu, M. Bismuth, and A. Aspuru-Guzik, "Computer vision for recognition of materials and vessels in chemistry lab settings and the vector-labpics data set," *ACS central science*, pp. 1743–1752, 2020.

[5] Y. R. Wang, Y. Zhao, H. Xu, S. Eppel, A. Aspuru-Guzik, F. Shkurti, and A. Garg, "Mvtrans: Multi-view perception of transparent objects," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 3771–3778.

[6] Z. Wu, S. Su, Q. Chen, and R. Fan, "Transparent cbjects: A corner case in stereo matching," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 12 353–12 359.

[7] J. Kerr, L. Fu, H. Huang, Y. Avigal, M. Tancik, J. Ichnowski, A. Kanazawa, and K. Goldberg, "Evo-NeRF: Evolving NeRF for sequential robot grasping of transparent objects," in *Conference on Robot Learning (CoRL)*, 2022, pp. 353–367.

[8] K. Tanaka, Y. Mukaigawa, H. Kubo, Y. Matsushita, and Y. Yagi, "Recovering transparent shape from time-of-flight distortion," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 4387–4395.

[9] 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.

[10] S. Sajjan, M. Moore, M. Pan, G. Nagaraja, J. Lee, A. Zeng, and S. Song, "Clear grasp: 3d shape estimation of transparent objects for manipulation," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2020, pp. 3634–3642.

[11] H. Xu, Y. R. Wang, S. Eppel, A. Aspuru-Guzik, F. Shkurti, and A. Garg, "Seeing glass: Joint point-cloud and depth completion for transparent objects," in *Conference on Robot Learning (CoRL)*, 2021, pp. 827–838.

[12] T. Tang, J. Liu, J. Zhang, H. Fu, W. Xu, and C. Lu, "Rftrans: Leveraging refractive flow of rransparent objects for surface normal estimation and manipulation," *IEEE Robotics and Automation Letters*, vol. 9, no. 4, pp. 3735–3742, 2024.

[13] L. Li, S. Khan, and N. Barnes, "Silhouette-assisted 3d object instance reconstruction from a cluttered scene," in *IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 2019, pp. 2080–2088.

[14] A. J. Perez, J. Perez-Soler, J.-C. Perez-Cortes, and J.-L. Guardiola, "Alignment and improvement of shape-from-silhouette reconstructed 3d objects," *IEEE Access*, vol. 12, pp. 76 975–76 985, 2024.

[15] K. N. Kutulakos and S. M. Seitz, "A theory of shape by space carving," *International Journal of Computer Vision*, vol. 38, pp. 199–218, 2000.

[16] X. Zuo, C. Du, S. Wang, J. Zheng, and R. Yang, "Interactive visual hull refinement for specular and transparent object surface reconstruction," in *IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 2237–2245.

[17] M. Piccardi, "Background subtraction techniques: a review," in *IEEE International Conference on Systems, Man and Cybernetics (SMC)*, vol. 4, 2004, pp. 3099–3104.

[18] J. Canny, "A computational approach to edge detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, no. 6, pp. 679–698, 1986.

[19] F. Qi, X. Tan, Z. Zhang, M. Chen, Y. Xie, and L. Ma, "Glass makes blurs: Learning the visual blurriness for glass surface detection," *IEEE Transactions on Industrial Informatics*, vol. 20, no. 4, pp. 6631–6641, 2024.

[20] F. Liu, Y. Liu, J. Lin, K. Xu, and R. W. Lau, "Multi-view dynamic reflection prior for video glass surface detection," in *AAAI Conference on Artificial Intelligence (AAAI)*, vol. 38, no. 4, 2024, pp. 3594–3602.

[21] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, *et al.*, "Segment anything," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 4015–4026.

[22] W. Matusik, C. Buehler, R. Raskar, S. J. Gortler, and L. McMillan, "Image-based visual hulls," in *Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, 2000, pp. 369–374.

[23] J.-S. Franco and E. Boyer, "Fusion of multiview silhouette cues using a space occupancy grid," in *IEEE International Conference on Computer Vision (ICCV)*, vol. 2, 2005, pp. 1747–1753.

[24] J. J. Park, P. Florence, J. Straub, R. Newcombe, and S. Lovegrove, "Deepsdf: Learning continuous signed distance functions for shape representation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 165–174.

[25] L. Mescheder, M. Oechsle, M. Niemeyer, S. Nowozin, and A. Geiger, "Occupancy networks: Learning 3d reconstruction in function space," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 4460–4470.

[26] B. Wen, J. Tremblay, V. Blukis, S. Tyree, T. Müller, A. Evans, D. Fox, J. Kautz, and S. Birchfield, "Bundlesdf: Neural 6-dof tracking and 3d reconstruction of unknown objects," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 606–617.

[27] E. Remelli, A. Lukoianov, S. Richter, B. Guillard, T. Bagautdinov, P. Baque, and P. Fua, "Meshsdf: Differentiable iso-surface extraction," *Advances in Neural Information Processing Systems*, vol. 33, pp. 22 468–22 478, 2020.

[28] Z. Chen, A. Tagliasacchi, and H. Zhang, "Bsp-net: Generating compact meshes via binary space partitioning," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 45–54.

[29] T. Shen, J. Gao, K. Yin, M.-Y. Liu, and S. Fidler, "Deep marching tetrahedra: a hybrid representation for high-resolution 3d shape synthesis," *Advances in Neural Information Processing Systems*, vol. 34, pp. 6087–6101, 2021.

[30] G. M. Treece, R. W. Prager, and A. H. Gee, "Regularised marching tetrahedra: improved iso-surface extraction," *Computers & Graphics*, vol. 23, no. 4, pp. 583–598, 1999.

[31] M. Ji, R.-Z. Qiu, X. Zou, and X. Wang, "Graspsplats: Efficient manipulation with 3d feature splatting," in *Conference on Robot Learning (CoRL)*, 2024, pp. 1443–1460.

[32] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," in *International Conference on Neural Information Processing Systems (NIPS)*, 2017, p. 5105–5114.

[33] H.-S. Fang, C. Wang, M. Gou, and C. Lu, "Graspnet-1billion: A large-scale benchmark for general object grasping," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 11 444–11 453.

[34] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, Q. Jiang, C. Li, J. Yang, H. Su, *et al.*, "Grounding dino: Marrying dino with grounded pre-training for open-set object detection," in *European Conference on Computer Vision (ECCV)*, 2024, pp. 38–55.

[35] A. Gropp, L. Yariv, N. Haim, M. Atzmon, and Y. Lipman, "Implicit geometric regularization for learning shapes," in *Conference on Machine Learning and Systems (MLSys)*, 2020, pp. 3569–3579.

[36] H.-S. Fang, C. Wang, H. Fang, M. Gou, J. Liu, H. Yan, W. Liu, Y. Xie, and C. Lu, "Anygrasp: Robust and efficient grasp perception in spatial and temporal domains," *IEEE Transactions on Robotics*, vol. 39, no. 5, pp. 3929–3945, 2023.

[37] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3d gaussian splatting for real-time radiance field rendering," *ACM Transactions on Graphics*, vol. 42, no. 4, pp. 139:1–139:14, 2023.

[38] J. Liu, H. Ma, Y. Guo, Y. Zhao, C. Zhang, W. Sui, and W. Zou, "Monocular depth estimation and segmentation for transparent object with iterative semantic and geometric fusion," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2025, pp. 11 162–11 168.