2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

ClearDepth: Efficient Stereo Perception of Transparent Objects for Robotic Manipulation

Kaixin Bai[1,2], Huajian Zeng[2,3,4], Lei Zhang[1,2†], Yiwen Liu[2,3], Hongli Xu[3], Zhaopeng Chen[2], Jianwei Zhang[1]

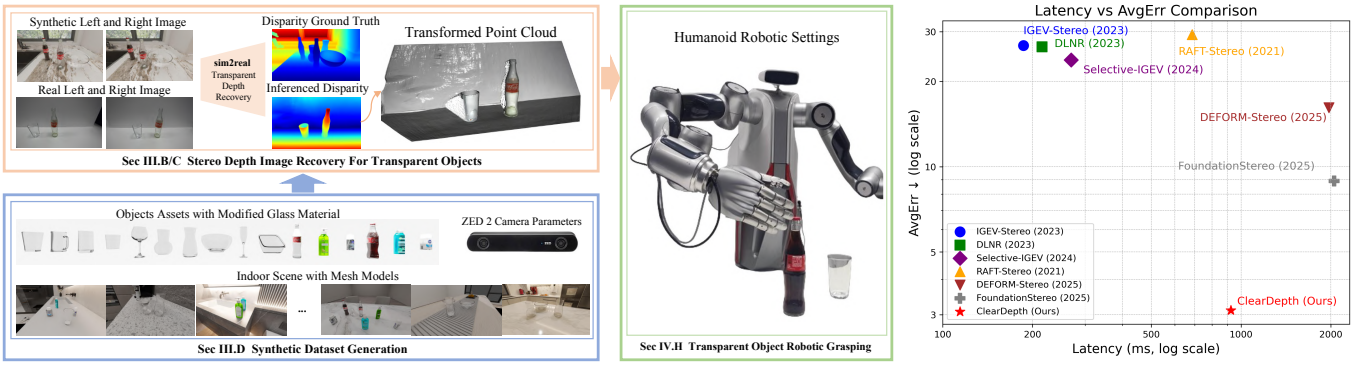


Fig. 1: ClearDepth leverages structure-aware stereo matching and synthetic training data to bridge the Sim2Real gap in transparent object grasping, achieving superior speed–accuracy trade-offs.

***Abstract*—Transparent object depth perception remains a major challenge in robotics and logistics due to the limitations of standard 3D sensors in capturing accurate depth on transparent and reflective surfaces. This affects applications relying on depth maps and point clouds, particularly in robotic manipulation. To address this, we propose ClearDepth, a vision transformer-based algorithm for stereo depth recovery of transparent objects, enhanced by a novel feature post-fusion module that refines depth estimation using structural visual features. To mitigate the high costs of stereo dataset collection, we introduce a physically realistic, domain-adaptive Sim2Real framework for efficient data generation. Our method outperforms state-of-the-art stereo matching approaches on transparent depth recovery. Furthermore, in transparent object grasping experiments, ClearDepth improves transparent-scene perception and achieves at least an 18% higher grasp success rate compared to the state-of-the-art methods for transparent object manipulation. Our method demonstrates strong Sim2Real generalization, enabling precise depth perception of transparent objects for robotic applications in the real world. Dataset and project details are available at https://sites.google.com/view/cleardepth/.**

I. INTRODUCTION

Transparent objects, such as glass bottles and cups, are prevalent in domestic service robotics and logistics sorting scenarios. However, their inherent transparency, particularly the complex effects of refraction and reflection, poses significant challenges for visual perception and recognition [1]. These perception limitations, in turn, constrain the robot's ability to manipulate such objects effectively in real-world tasks.

Deep learning has played a critical role in understanding and modeling the complex geometrical features of transparent objects. To address these challenges, prior research has primarily focused on enhancing perception capabilities through deep learning, such as reconstructing depth from incomplete depth maps [2], [3], stereo visual perception [4], and multi-view approaches [5]. Despite notable progress, real-world applications still face difficulties in extracting reliable feature points due to inconsistencies in depth data input and the increased complexity of multi-view imaging systems. Transparent objects refract background textures, making structural features more critical than texture features for imaging and perception. To obtain stable feature points, some studies have explored extracting structural details [6] or improving the precision of depth sensing hardware [7]. However, these approaches remain limited in effectiveness and generalization ability.

Studies have shown that CNNs excel at texture recognition, while vision Transformers (ViTs) demonstrate superior capabilities in modeling shape features [8]. However, traditional ViTs typically downsample the input and rely on learnable upsampling to restore spatial resolution. While effective, this approach is computationally expensive and often lacks the ability to capture fine-grained details. RAFT-Stereo [9], an extension of RAFT [10], applies optical flow techniques to stereo matching, improving generalization and robustness with its lightweight Gated Recurrent Unit (GRU) Network module, but struggles with global context extraction due to its CNN architecture. To address these limitations, models such as SegFormer [11] and DinoV2 [12] enhance ViTs through cascaded architectures and multi-scale feature fusion, improving performance in depth estimation and semantic segmentation tasks. Transparent objects pose additional challenges due to their optical properties, which often cause background textures to become distorted. As a result, texture-based features become unreliable, and structural features become crucial for accurate perception. To better capture these structural cues, we design an efficient **cascaded ViT backbone** to extract contextual structural information, making it well-suited for modeling transparent object scenes.

†Corresponding author. zhanglei.cn.de@gmail.com,lei.zhang-1@studium.uni-hamburg.de
[1]TAMS (Technical Aspects of Multimodal Systems), Department of Informatics, University of Hamburg, Hamburg, Germany.
[2]Agile Robots SE, Munich, Germany.
[3]Technical University of Munich, Germany.
[4]Mohamed Bin Zayed University of Artificial Intelligence (MBUZAI), Abu Dhabi, UAE.

Moreover, conventional stereo matching networks typically rely on dot-product similarity for feature correspondence, which is ineffective in transparent object scenarios due to background refraction. To overcome this, we introduce a lightweight **post-fusion module** that incorporates structural feature priors into the Gated Recurrent Unit (GRU) update loop. This design improves structural awareness without introducing the computational overhead of cross-attention mechanisms. The whole pipeline is shown in Fig. 1.

Accurate datasets are vital for deep learning on transparent objects, yet existing collection methods, such as pose markers [13], [14], opaque substitutes [15], and manual 3D modeling [16], are labor-intensive and yield noisy depth maps. To address this, simulation engines are increasingly used [5], [16], though balancing realism and efficiency remains a challenge. We propose **SynClearDepth**, a synthetic dataset generated via a realistic data generation pipeline that supports direct model deployment on real-world sensors, providing instance segmentation, object poses, and depth maps.

In summary, our main contributions are:

1) An efficient stereo depth recovery network **ClearDepth** for transparent objects, featuring a cascaded ViT encoder for multi-scale structural feature extraction and a lightweight post-fusion module that integrates structural priors with appearance cues to achieve robust and efficient depth estimation.
2) The demonstrated advancements over SOTA methods, as evidenced in stereo perception benchmarks and real-world scenarios, exhibit significant qualitative and quantitative enhancements in the robotic grasping of transparent objects in single-object and cluttered environments, underscoring our solution's superior effectiveness.
3) SynClearDepth, a photo-realistic dataset for transparent object perception in grasping scenes, containing 14,091 stereo RGB images with ground-truth depth and segmentation labels. It aligns simulated with real sensor parameters and leverages domain randomization and adaptation to ensure diversity and robustness across different scenes and camera settings.

II. Related Work

A. Transparent Object Perception

Robotic perception of transparent objects remains challenging due to their low contrast and complex light interactions, which affect sensor accuracy in determining position and shape. Traditional RGB and RGB-D cameras struggle with these objects, as they rely on intensity data and overlook optical properties. To address this, research has explored polarized cameras, which reduce reflections and enhance contrast [7], [17]. However, their high cost limits widespread adoption. Alternative approaches include CNN-transformer-based models for tracking [18], Sim2Real techniques leveraging synthetic datasets [19], and alpha-matting methods for transparent object segmentation [20]. For robotic manipulation and pose estimation, multi-task perception models have been introduced [13], [16], [21]. Depth recovery remains particularly challenging due to light refraction and reflection. Methods such as NeRF and volumetric rendering aid surface reconstruction [22]–[24], while stereo and multi-view techniques improve depth estimation [4], [14], [25]. These approaches leverage various sensors, including RGB-D, stereo vision, and multi-view systems, to enhance transparent object perception [2]–[5], [7], [13], [14], [16], [21], [23]. Advances in deep learning and sensor technologies continue to drive improvements in accuracy and reliability.

B. Deep Learning-based Stereo Depth Recovery

Deep learning-based stereo matching methods have recently outperformed traditional approaches, with 2D convolutional models [26], [27] offering simplicity and efficiency. These models achieve high accuracy even on limited computational resources, making them suitable for engineering applications, though they still require improvements in accuracy and robustness due to 3D cost space constraints. 3D convolutional networks [28], [29] provide better interpretability and higher disparity map accuracy but require optimization due to their computational demands. STTR [30], inspired by SuperGlue [31], uses transformers with positional embedding and attention mechanisms for binocular dense matching, producing disparity and depth maps. However, these methods are computationally intensive and slow in inference, limiting their suitability for high-resolution images and downstream robotic tasks.

C. Transparent Object Datasets

Recent works [32], [33] show that real-world datasets degrade model performance due to label noise, while synthetic data with precise labels, enhance model performance. However, synthetic datasets across different data domains remain scarce in the open-source community. Ray-tracing renderers have narrowed the sim2real gap, making domain differences the main bottleneck in model generalization. Existing synthetic datasets, such as [14], [15], [21], [34]–[36], typically feature transparent objects on desktops. However, these datasets lack the complexity necessary for generalization to real-world scenarios like kitchens, bedrooms, and offices, where service and humanoid robots operate. Moreover, these datasets often require extensive pre- or post-processing, such as segmentation or background reconstruction, which is impractical for end-to-end algorithms crucial to embodied intelligence applications. Datasets using HDRI backgrounds [5] lack depth labels, which hinders generalization, especially in zero-shot tasks. Others simulate Realsense cameras [21], [36], but their rendering pipelines are complex and inefficient. To address these gaps, our dataset provides richly annotated indoor scenes with realistic transparent objects (e.g., containers, cosmetics), complete with background depth maps. It is designed to support scalable, efficient training for future embodied intelligence applications.

III. Problem Statement and Methods

A. Problem Statement

Stereo depth estimation for transparent objects is fundamentally challenging because the observed pixel intensity is not solely determined by the surface geometry but also influenced by background refraction and reflection. In other words, the imaging process of transparent objects often mixes optical information from the background, making traditional appearance- or texture-based stereo matching inherently ill-posed.

Formally, the intensity $I(x)$ at pixel x can be expressed as:

$$I(x) = \alpha\, T(x) + (1 - \alpha)\, B(x), \tag{1}$$

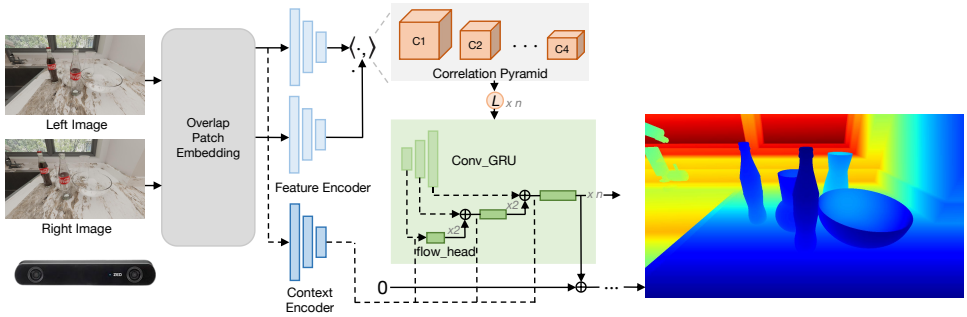


Fig. 2: Our stereo depth recovery network for transparent objects. The feature encoder extracts appearance features from both left and right images, while a context encoder processes the left image to provide structural priors for disparity refinement. A correlation pyramid is then constructed by merging left–right features to capture correspondence cues. These features, together with structural priors, are iteratively refined through a GRU-based update loop, which integrates texture similarity and structural consistency. The network finally outputs a refined disparity map that is robust to transparency-induced ambiguities.

where $T(x)$ denotes the transmitted (refracted) signal, $B(x)$ represents the background contribution, and $\alpha \in [0, 1]$ is the transparency coefficient.

To address this limitation, ClearDepth incorporates **structural features** that are more invariant to transparency effects. Given a disparity field $d(x)$, the stereo matching objective can be formulated as:

$$\min_{d(x)} \|I_L(x) - I_R(x - d(x))\|^2 + \lambda \mathcal{R}(d(x), \phi_s(x)), \quad (2)$$

where I_L, I_R denote the stereo image pair, and $\mathcal{R}(d(x), \phi_s(x))$ is a structural regularizer that enforces consistency between disparity and structural embeddings $\phi_s(x)$. These embeddings are extracted via a cascaded ViT backbone, whose global self-attention mechanism captures **long-range shape and contour information**, thereby reducing ambiguities in transparent regions. The cascaded vision transformer backbone is detailed in Sec. III-B. Furthermore, to compensate for the failure modes of traditional dot-product similarity, we also introduce a **post-fusion mechanism** that combines texture-based and structure-based disparity estimates:

$$d_f(x) = w_s(x)\, d_s(x) + w_a(x)\, d_a(x), \quad w_s(x) + w_a(x) = 1, \quad (3)$$

where $d_a(x)$ is the disparity derived from appearance similarity, $d_s(x)$ is the structure-guided disparity, and $w_s(x), w_a(x)$ are adaptive confidence weights. Intuitively, when texture cues are reliable (opaque regions), $w_a(x)$ dominates, while in transparent or textureless regions, $w_s(x)$ dominates, enforcing structural consistency. The fusion design is introduced in Sec. III-C.

In summary, we propose **ClearDepth**, employ a ViT backbone for robust structural feature extraction and design a post-fusion module to explicitly compensate for the shortcomings of traditional stereo matching in transparent-object scenarios. Our network is illustrated in Fig. 2. The dataset generation is presented in Sec. III-D.

B. Cascaded Vision Transformer Backbone.

Our backbone begins with overlap patch embedding for initial tokenization, preserving local features. Tokens pass through four transformer blocks, generating feature maps at $\frac{1}{4}$, $\frac{1}{8}$, $\frac{1}{16}$, and $\frac{1}{32}$ scales. To optimize computational efficiency, the model incorporates efficient self-attention, which significantly reduces the computational burden from $O(N^2)$ to $O(\frac{N^2}{R})$. This reduction is achieved by first reshaping the input sequence from $N \cdot C$ to $\frac{N}{R} \times (C \cdot R)$ by 2d convolutional layer with the stride $8, 4, 2, 1$ for different ViT blocks, as detailed in Equ. 4, and then adjusting the sequence dimensions back to C channel through linear layers, as described in Equ. 5. K denotes the sequence in the ViT block optimized for lower computational complexity.

$$\hat{K} = Reshape(\frac{N}{R}, C \cdot R)(K) \quad (4)$$

$$K = Linear(C \cdot R, C)(\hat{K}) \quad (5)$$

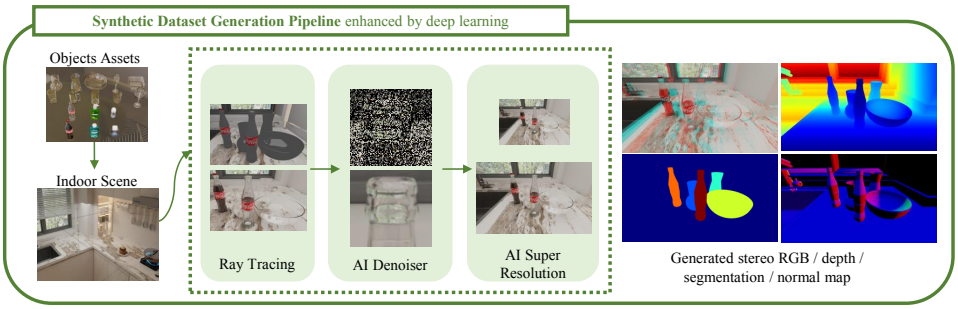
Additionally, the Mix-FFN module in the architecture addresses the challenge of performance degradation due to the interpolation of positional embeddings in the original ViT structure, especially when dealing with varying input image sizes, here by substituting positional embeddings with learnable depth-wise convolutions. The equation is as 6.

$$\mathbf{x}_{\text{out}} = \text{MLP}(\text{GELU}(\text{Conv}_{3\times3}(\text{MLP}(\mathbf{x}_{\text{in}})))) + \mathbf{x}_{\text{in}} \quad (6)$$

Then, we concatenate multi-scale feature maps from different ViT blocks by upsampling them to a unified scale of $\frac{1}{4}$. This combined feature map undergoes further refinement through a precise $1 \cdot 1$ convolution, facilitating optimal dimension adjustment.

C. Structural Feature Post-Fusion

We propose a modified GRU-based architecture that refines disparity maps in a coarse-to-fine manner. The Post-Fusion mechanism is specifically designed to address the unique challenges of transparent objects. Our experiments show that, unlike opaque objects, accurate depth estimation of transparent objects relies heavily on fine-grained structural cues. In addition, the refractive nature of transparent surfaces distorts background textures, making dot-product–based feature similarity unreliable for reconstruction. To mitigate this, we incorporate structural information from the image itself into the GRU iterations. This integration ensures that structural cues extracted at multiple resolutions are consistently preserved throughout the iterative



(a) Pipeline of synthetic dataset generation.

(b) Rendered 3D models and real-world objects

Fig. 3: SynClearDepth dataset with diverse objects, various scene configurations.

refinement process.

The core update equations in our model are defined as follows:

$$x_k = [\mathbf{C}_k, \mathbf{d}_k, \mathbf{c}_k, \mathbf{c}_r, \mathbf{c}_h] \tag{7}$$

$$z_k = \sigma(\text{Conv}([h_{k-1}, x_k], W_z) + c_k), \tag{8}$$

$$r_k = \sigma(\text{Conv}([h_{k-1}, x_k], W_r) + c_r), \tag{9}$$

$$\tilde{h}_k = \tanh(\text{Conv}([r_k \odot h_{k-1}, x_k], W_h) + c_h), \tag{10}$$

$$h_k = (1 - z_k) \odot h_{k-1} + z_k \odot \tilde{h}_k, \tag{11}$$

Here, x_k is a concatenation of several feature maps, including the correlation $\mathbf{C}_k$, the current disparity $\mathbf{d}_k$, and structural context feature maps $\mathbf{c}_k$, $\mathbf{c}_r$, and $\mathbf{c}_h$. Specifically, $\mathbf{c}_k$, $\mathbf{c}_r$, and $\mathbf{c}_h$ represent structural features derived from the left image. These features are incorporated as residuals into the GRU loop, allowing for enhanced participation of structural information during the disparity map refinement process. z, r, h represent the state information of the update gate, reset gate, and hidden gate in a GRU.

Then, Our approach decode GRUs at each resolutions to obtain multi-scale disparity updates for coarse to fine gradual optimization:

$$\triangle \mathbf{d}_{k,\frac{1}{32}} = \text{Decoder}(h_{k,\frac{1}{32}}), \tag{12}$$

$$\triangle \mathbf{d}_{k,\frac{1}{16}} = \text{Decoder}(h_{k,\frac{1}{16}} + \text{Interp}(\triangle \mathbf{d}_{k,\frac{1}{32}})), \tag{13}$$

$$\triangle \mathbf{d}_{k,\frac{1}{8}} = \text{Decoder}(h_{k,\frac{1}{8}} + \text{Interp}(\triangle \mathbf{d}_{k,\frac{1}{16}})), \tag{14}$$

where Decoder consist of two convolutional layers and Interp is bilinear interpolation scaled up by a factor of two. Finally, the updated disparity is calculated as:

$$\mathbf{d}_{k+1} = \mathbf{d}_k + \triangle \mathbf{d}_k \tag{15}$$

In summary, to address the challenges of transparent objects, we selected an appropriate image feature extractor. Additionally, considering the unique difficulties of transparent objects and the need for efficient models in robotics, we designed a structural feature post-fusion architecture. Every detail of our network structure is tailored to the characteristics of transparent object scenarios.

In the comparative experiments section, the visual results demonstrate that our model substantially enhances the stereo imaging of transparent objects.

D. Synthetic Dataset Generation

To enhance the efficiency of synthetic dataset generation, we utilized the AI denoiser provided by OptiX [37] during rendering and adopted open-source pretrained deep learning super-resolution [38] as rendering output optimization strategies, reducing the average generation time per set (stereo RGB, depth, masks, and object-camera poses) from 12.77 to 4.40 seconds. Since these techniques are widely used in computer graphics and do not alter the core data distribution, we omit further analysis. The dataset generation process is illustrated in Fig. 3a. Our SynClearDepth dataset includes 16 selected objects: 10 common transparent containers and 6 glass-material products (Fig. 3b). To ensure depth labels for both objects and backgrounds, we combined object models with indoor scenes, including 6 bathrooms, 3 dining rooms, 5 kitchens, and 6 living rooms. This resulted in 14,091 image sets, each containing left and right RGB images, ground truth depth, instance segmentation, and object/camera poses (Fig. **??**). We applied domain randomization to object types, quantities, poses, lighting, and camera angles. This dataset is designed to support robotic perception and manipulation in service robot applications, particularly for handling transparent objects in household environments. For more details, please refer to the supplementary materials.

IV. Experiments

A. Technical Specifications

Our network is firstly pre-trained on CREStereo dataset [39] and Scene Flow dataset [26], and then fine-tuned on our proposed SynClearDepth dataset for transparent object stereo imaging. Our model is trained on 1 block of NVIDIA RTX A6000 with batch size 8 and the whole training lasts for 300,000 steps. We use AdamW [40] as optimizer, the learning rate is set to 0.0002, updated with a warm-up mechanism and used one-cycle learning rate scheduler. The final learning rate when training finished is 0.0001. The input size of the model is resized to 360×720. Fine-tune for transparent objects takes the same training parameters as pretraining on the opaque dataset.

B. Evaluation Metrics

1) **AvgErr (Average Error):** Represents the average disparity error across all pixels, indicating the general accuracy of the disparity map.

2) **RMS (Root Mean Square Error):** Measures the square root of the average squared disparity error, reflecting the overall deviation from the ground truth.
3) **Bad 0.5 (%), Bad 1.0 (%), Bad 2.0 (%), Bad 4.0 (%):** These metrics indicate the percentage of pixels where the disparity error exceeds 0.5, 1.0, 2.0, and 4.0 pixels, respectively, highlighting the proportion of significant errors in the disparity map.

Together, these metrics provide a comprehensive assessment of stereo matching performance, balancing both overall accuracy and the frequency of large errors.

TABLE I: Quantitative results on transparent object dataset compared with stereo SOTA methods fine-tuned with SynClearDepth dataset. Visualization results shown in Fig. 4.

Methods	AvgErr ↓	RMS ↓	bad 0.5 (%) ↓	bad 1.0 (%) ↓	bad 2.0 (%) ↓	bad 4.0 (%) ↓
IGEV-Stereo [41]	2.077	5.5301	58.9743	36.1270	19.967	10.777
DLNR [42]	3.097	8.4269	28.1088	21.8442	16.481	11.9046
Selective-IGEV [43]	**1.273**	**4.3365**	34.8229	17.6288	**9.561**	5.8707
RAFT-Stereo [9]	2.245	8.8016	29.7356	17.4521	10.835	6.2107
ClearDepth (ours)	2.138	8.7282	**24.7329**	**16.3178**	9.8459	**5.7600**

C. Qualitative and Quantitative Studies for Stereo Depth Estimation

1) Evaluation on Transparent Object Dataset: To validate our model and dataset for transparent object depth recovery in stereo vision, we fine-tuned our pre-trained model on the SynClearDepth dataset using the same training parameters as pre-training. We also fine-tuned RAFT-Stereo [9], IGEV-Stereo [41], DLNR [42], Selective-IGEV [43] from the Middlebury benchmark on SynClearDepth for comparison. This analysis highlights our model's improvements in stereo-based depth perception for transparent objects. Tab. I presents quantitative results, while Fig. 4 visualizes stereo imaging performance. AvgErr (Average Error) and RMS (Root Mean Square Error) measure numerical error, while Bad 0.5 (%), Bad 1.0 (%), Bad 2.0 (%), Bad 4.0 (%) reflect relative error. Results show our model achieves strong performance in numerical error and outperforms all others in relative error. Our model is more efficient than others, achieving comparable performance without the high computational cost of cross-attention or multi-model ensembles. This is due to innovations in the image encoder, making our approach more suitable for robotics.

2) Comparison experiments with SOTA zero-shot stereo matching methods: To evaluate the effectiveness of our method on transparent object stereo depth estimation, we conduct a comprehensive comparison against several SOTA open-source zero-shot stereo matching approaches [41]–[45] on a dedicated transparent-object validation set. Specifically, we include FoundationStereo [44], and DEFOM-Stereo [45], all of which claim to generalize to arbitrary unseen scenes without requiring additional training. For fair comparison, we directly adopt their officially released pretrained models and evaluate them under identical conditions involving transparent objects.

Given that our target application is robotic grasping, where the foreground regions (i.e., the object areas) are of primary importance, we compute all quantitative metrics exclusively on these regions to better reflect each model's performance on the most critical parts of the scene. As shown in Table II, our method substantially outperforms all competing methods across all evaluation metrics. It achieves lower average error, reduced root mean square (RMS) error, and the lowest bad-pixel rates under multiple threshold settings. These results demonstrate that our approach not only generalizes effectively to novel transparent-object scenes but also delivers substantial accuracy improvements over existing zero-shot stereo methods.

This also indicates that the lack of transparent-object stereo datasets in the current open-source community negatively impacts the performance of zero-shot stereo methods, further highlighting the value and contribution of our dataset.

TABLE II: Quantitative results on transparent object dataset compared with stereo SOTA zero-shot stereo reconstruction methods.

Methods	AvgErr ↓	RMS ↓	bad 0.5 (%) ↓	bad 1.0 (%) ↓	bad 2.0 (%) ↓	bad 4.0 (%) ↓
IGEV-Stereo [41]	26.8047	39.0820	0.968369	0.939858	0.890477	0.783732
DLNR [42]	26.5240	38.0410	0.962301	0.927577	0.865248	0.758555
Selective-IGEV [43]	23.8168	36.2975	0.959417	0.921393	0.854298	0.731494
RAFT-Stereo [9]	29.2919	39.2978	0.971589	0.946811	0.901663	0.821213
DEFOM-Stereo [45]	16.1635	25.8188	0.890889	0.792165	0.680660	0.550519
FoundationStereo [44]	8.8985	16.0874	0.891110	0.799528	0.668938	0.498191
ClearDepth (ours)	**3.1084**	**6.9570**	**0.806759**	**0.631477**	**0.375651**	**0.153726**

TABLE III: Ablation study for the feature post-fusion module in clearDepth with 100,000 steps on SynClearDepth dataset.

Methods	AvgErr ↓	RMS ↓	bad 0.5 (%) ↓	bad 1.0 (%) ↓	bad 2.0 (%) ↓	bad 4.0 (%) ↓
w/o Fusion	6.90	15.48	43.34	29.63	21.52	16.62
Feature Fusion	**2.64**	**8.59**	**27.23**	**16.87**	**11.28**	**7.72**

3) Ablation Study of Feature Post-Fusion Module: To evaluate the impact of our feature post-fusion module, we conducted ablation studies on the SynClearDepth dataset. We compared networks with and without the module, as shown in Tab. III. Results indicate a substantial performance boost, especially in handling complex transparency and light refraction, highlighting its effectiveness in enhancing depth estimation and object recognition. Each study was trained for 100,000 steps.

4) Qualitative experiments on real-world scenes with different materials, lighting conditions: We perform qualitative analysis on real-world images with different materials and lighting conditions, as shown in Fig. 5 and Fig. 6. We compare depth perception performance for objects with different materials using our method with SOTA methods, including ClearGrasp [15], TransCG [13], ASGrasp [36]. For more results, please check out our supplementary materials and videos. Results in supplementary video show that leveraging a physically realistic renderer enables strong generalization in real world, with performance consistent across domain-shifted test sets. In this work, we adopt a stereo-based approach instead of the Realsense-based imaging methods [36]. The open-source datasets using Realsense cameras are limited in both diversity and scale compared to stereo datasets, restricting future extensions. Additionally, Realsense IR projection penetrates transparent objects, leading to visual information loss [36], which we avoid to ensure robustness.

D. Trade-off between Speed and Accuracy

1) Comparison experiment of inference speed and FLOPs: We compare the speed and average error of our method and

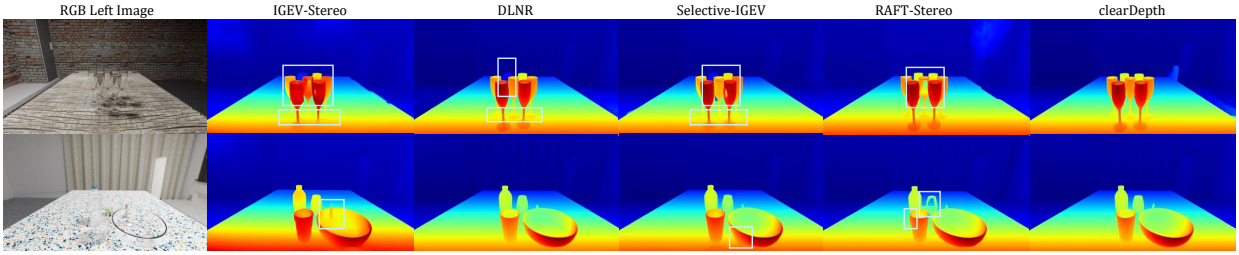


Fig. 4: The visualization results of our transparent object stereo depth reconstruction method compare with other SOTA stereo depth estimation methods by fine-tuning on SynClearDepth dataset.

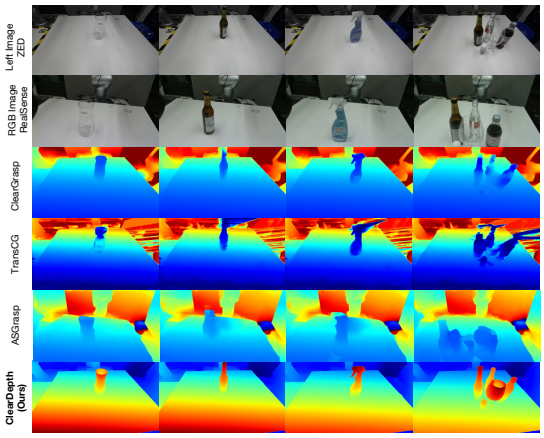


Fig. 5: Qualitative experiments of ClearGrasp [15], TransCG [13], ASGrasp [36] and proposed ClearDepth for objects with different materials in single-object and cluttered scene.

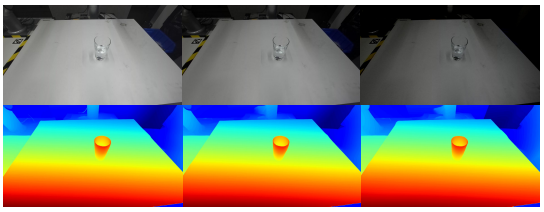


Fig. 6: Qualitative experiments of proposed ClearDepth for scenes with different lighting conditions.

SOTA methods, as shown in Fig. 1. For detailed data, please refer to the supplementary material.

2) TensorRT implementation: Additionally, our TensorRT implementation enables real-time inference at 50 FPS on consumer GPUs, whereas other models, due to their complex designs, are impractical for deployment.

TABLE IV: Real-world robotic grasping comparison experiments with SOTA methods for transparent objects.

Grasp SR	single (L1)	cluttered (L1)	single (L2)	cluttered (L2)
Baseline [46]	78%	63%	62%	58%
TransCG [13]	80%	70%	78%	67%
ClearDepth (ours)	98%	92%	98%	90%

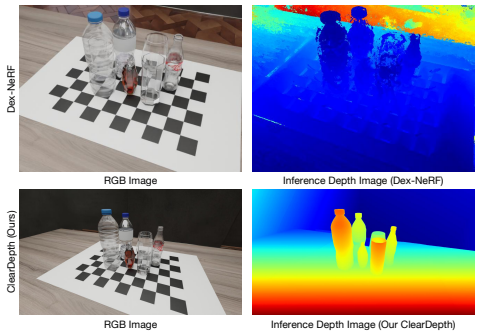


Fig. 7: Comparison experiment with NeRF-based methods [35].

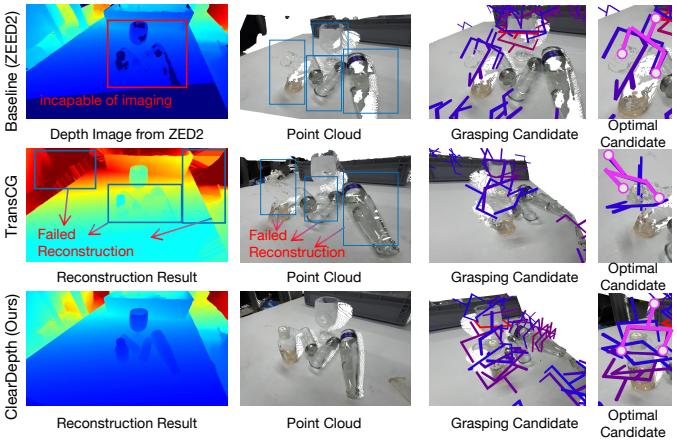


Fig. 8: Real-world qualitative comparisons of transparent object grasping using depth reconstruction of ZED2 [46], TransCG [13], our ClearDepth. The grasping candidates are estimated using GraspNet-Baseline [47]. Depth images, point clouds, and grasping results are presented.

E. Comparison experiment with NeRF-based method

We execute comparison experiment with NeRF-based method [35]. The reconstructed depth images are shown in Fig. 7. Our method achieves better reconstruction quality compared to NeRF-based method [35], which requires additional data acquisition and suffers from low efficiency. In the context of robotic manipulation tasks, training a separate model for each scene introduces considerable overhead.

F. Additional experiments

We execute addifional comparison experiments with [9], [39], [48] in Middleburry dataset [49] and KITTI dataset, as detailed in Supplementary Materials.

G. Comparison Experiments of Transparent Object Grasping

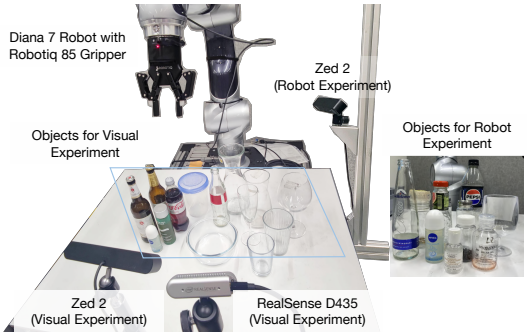


Fig. 9: Experiment setup for grasping comparison experiment.

To evaluate the performance of our transparent object grasping pipeline compared to state-of-the-art (SOTA) methods [13], we conducted real-world experiments involving two-finger grasps on transparent objects, as shown in Fig. 9. The depth data based on the stereo reconstruction method [46] from the ZED camera is used for grasp generation as a baseline method. The evaluation scenarios include both single-object grasping and grasping in cluttered environments. Specifically, Level-1 (L1) scenes contain a mix of transparent and opaque objects, while Level-2 (L2) scenes consist exclusively of fully transparent objects. For each experimental setting, we performed 150 grasping trials. The grasp success rate is computed as the number of successful grasps divided by the total number of attempts. The grasp success rates for all methods are summarized in the Tab. IV, and the corresponding reconstruction results and grasp predictions are illustrated in the Fig. 8. Our method consistently achieves the highest performance across all levels of scene and object complexity. Specifically, it demonstrates superior grasp success rates in both single-object and multi-object scenarios.

1) Analysis of Robotic Grasping Experiments: To evaluate the effectiveness of our method, we conducted an in-depth analysis of the causes of grasp failures. The primary cause of failure lies in inaccurate depth reconstruction, which directly leads to unsuccessful grasp attempts. Additionally, grasp prediction errors may also result in collisions or object drops during execution. Specifically, the limitations of depth reconstruction manifest in two ways: (1) the inability to perceive transparent regions, leading to collisions between the gripper and the object during execution; and (2) the prediction of noisy points within transparent regions, causing grasp candidates to be located in unreliable areas, ultimately resulting in failure. The distribution of failure causes across different methods is shown in Fig. 10. It is evident that our method substantially reduces the proportion of failures caused by inaccurate depth reconstruction to increase the grasping success rate.

H. Multi-Fingered Robotic Grasping Experiment

We also employ our pipeline for transparent object grasping in a robot platform with a robotic arm and multi-fingered robotic hand, as shown in Fig. 1. Following grasping pipeline from ContactDexNet [50], multi-fingered robotic grasping experiment achieves an 86.2% average success rate.

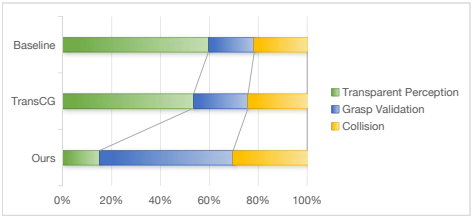


Fig. 10: Error distribution of baseline method, TransCG [13] and our ClearDepth. We compare the proportions of total failures represented by different failure types.

V. Conclusion and Future Work

In this work, we present a complete visual perception framework for transparent object manipulation in service robotics scenarios, spanning synthetic data generation, stereo depth estimation, and real-world robotic validation. We propose an efficient real-time stereo depth recovery network that combines a cascaded vision transformer backbone with a structural feature post-fusion module, enabling fine-grained structural perception and accurate depth recovery of transparent objects without relying on mask priors. To address the data scarcity challenge in transparent object perception, we construct SynClearDepth, a high-quality simulation dataset containing diverse household environments and realistic object placements. It provides accurate RGB, depth maps, instance masks, and pose annotations, substantially enhancing model generalization in real-world scenarios. We validate our model through extensive comparisons on public and proprietary datasets, along with ablation studies. Experimental results demonstrate that our approach outperforms existing methods on both public and proprietary benchmarks, particularly in structure-aware and boundary-level depth estimation. Results demonstrate its robustness, accuracy, and efficiency, supporting transparent object manipulation in robotics. Furthermore, real-world robotic grasping experiments show that our method can be seamlessly integrated into grasping pipelines without requiring multi-view capture or additional pre-processing, and achieves stable and precise manipulation of transparent objects. These results highlight the practicality and applicability of stereo-based transparent object depth estimation in real-world robotic tasks.

References

[1] J. Jiang, G. Cao, J. Deng, T.-T. Do, and S. Luo, "Robotic perception of transparent objects: A review," *IEEE Transactions on Artificial Intelligence*, 2023. 1

[2] T. Li, Z. Chen, H. Liu, and C. Wang, "Fdct: Fast depth completion for transparent objects," *IEEE Robotics and Automation Letters*, 2023. 1, 2

[3] K. Chen, S. Wang, B. Xia, D. Li, Z. Kan, and B. Li, "Tode-trans: Transparent object depth estimation with transformer," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 4880–4886. 1, 2

[4] K. Chen, S. James, C. Sui, Y.-H. Liu, P. Abbeel, and Q. Dou, "Stereopose: Category-level 6d transparent object pose estimation from stereo images via back-view nocs," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 2855–2861. 1, 2

[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 *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 3771–3778. 1, 2

[6] Y. Cao, Z. Zhang, E. Xie, Q. Hou, K. Zhao, X. Luo, and J. Tuo, "Fakemix augmentation improves transparent object detection," *arXiv preprint arXiv:2103.13279*, 2021. 1

[7] M. Shao, C. Xia, D. Duan, and X. Wang, "Polarimetric inverse rendering for transparent shapes reconstruction," *arXiv preprint arXiv:2208.11836*, 2022. 1, 2

[8] S. Tuli, I. Dasgupta, E. Grant, and T. L. Griffiths, "Are convolutional neural networks or transformers more like human vision?" *arXiv preprint arXiv:2105.07197*, 2021. 1

[9] L. Lipson, Z. Teed, and J. Deng, "Raft-stereo: Multilevel recurrent field transforms for stereo matching," in *2021 International Conference on 3D Vision (3DV)*. IEEE, 2021, pp. 218–227. 1, 5, 7

[10] Z. Teed and J. Deng, "Raft: Recurrent all-pairs field transforms for optical flow," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*. Springer, 2020, pp. 402–419. 1

[11] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," *Advances in Neural Information Processing Systems*, vol. 34, pp. 12 077–12 090, 2021. 1

[12] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby *et al.*, "Dinov2: Learning robust visual features without supervision," *arXiv preprint arXiv:2304.07193*, 2023. 1

[13] 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. 2, 5, 6, 7

[14] 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," *arXiv preprint arXiv:2110.00087*, 2021. 2

[15] 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 *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 3634–3642. 2, 5, 6

[16] X. Chen, H. Zhang, Z. Yu, A. Opipari, and O. Chadwicke Jenkins, "Clearpose: Large-scale transparent object dataset and benchmark," in *European Conference on Computer Vision*. Springer, 2022, pp. 381–396. 2

[17] A. Kalra, V. Taamazyan, S. K. Rao, K. Venkataraman, R. Raskar, and A. Kadambi, "Deep polarization cues for transparent object segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8602–8611. 2

[18] K. Garigapati, E. Blasch, J. Wei, and H. Ling, "Transparent object tracking with enhanced fusion module," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2023, pp. 7696–7703. 2

[19] A. Lukezic, Z. Trojer, J. Matas, and M. Kristan, "Trans2k: Unlocking the power of deep models for transparent object tracking," *arXiv preprint arXiv:2210.03436*, 2022. 2

[20] H. Cai, F. Xue, L. Xu, and L. Guo, "Transmatting: Tri-token equipped transformer model for image matting," *arXiv preprint arXiv:2303.06476*, 2023. 2

[21] Q. Dai, J. Zhang, Q. Li, T. Wu, H. Dong, Z. Liu, P. Tan, and H. Wang, "Domain randomization-enhanced depth simulation and restoration for perceiving and grasping specular and transparent objects," in *European Conference on Computer Vision*. Springer, 2022, pp. 374–391. 2

[22] Z. Li, X. Long, Y. Wang, T. Cao, W. Wang, F. Luo, and C. Xiao, "Neto: Neural reconstruction of transparent objects with self-occlusion aware refraction-tracing," *arXiv preprint arXiv:2303.11219*, 2023. 2

[23] Q. Dai, Y. Zhu, Y. Geng, C. Ruan, J. Zhang, and H. Wang, "Graspnerf: Multiview-based 6-dof grasp detection for transparent and specular objects using generalizable nerf," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 1757–1763. 2

[24] Z. Li, Y.-Y. Yeh, and M. Chandraker, "Through the looking glass: neural 3d reconstruction of transparent shapes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 1262–1271. 2

[25] H. Zhang, A. Opipari, X. Chen, J. Zhu, Z. Yu, and O. C. Jenkins, "Transnet: Transparent object manipulation through category-level pose estimation," *arXiv preprint arXiv:2307.12400*, 2023. 2

[26] N. Mayer, E. Ilg, P. Hausser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox, "A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4040–4048. 2, 4

[27] H. Xu and J. Zhang, "Aanet: Adaptive aggregation network for efficient stereo matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 1959–1968. 2

[28] Y. Cao, J. Xu, S. Lin, F. Wei, and H. Hu, "Gcnet: Non-local networks meet squeeze-excitation networks and beyond," in *Proceedings of the IEEE/CVF international conference on computer vision workshops*, 2019, pp. 0–0. 2

[29] J.-R. Chang and Y.-S. Chen, "Pyramid stereo matching network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5410–5418. 2

[30] Z. Li, X. Liu, N. Drenkow, A. Ding, F. X. Creighton, R. H. Taylor, and M. Unberath, "Revisiting stereo depth estimation from a sequence-to-sequence perspective with transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 6197–6206. 2

[31] P.-E. Sarlin, D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superglue: Learning feature matching with graph neural networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 4938–4947. 2

[32] L. Yang, B. Kang, Z. Huang, Z. Zhao, X. Xu, J. Feng, and H. Zhao, "Depth anything v2," *arXiv:2406.09414*, 2024. 2

[33] A. Bochkovskii, A. Delaunoy, H. Germain, M. Santos, Y. Zhou, S. R. Richter, and V. Koltun, "Depth pro: Sharp monocular metric depth in less than a second," *arXiv preprint arXiv:2410.02073*, 2024. 2

[34] L. Zhu, A. Mousavian, Y. Xiang, H. Mazhar, J. van Eenbergen, S. Debnath, and D. Fox, "Rgb-d local implicit function for depth completion of transparent objects," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 4649–4658. 2

[35] J. Ichnowski, Y. Avigal, J. Kerr, and K. Goldberg, "Dex-nerf: Using a neural radiance field to grasp transparent objects," *arXiv preprint arXiv:2110.14217*, 2021. 2, 6

[36] J. Shi, Y. Jin, D. Li, H. Niu, Z. Jin, H. Wang *et al.*, "Asgrasp: Generalizable transparent object reconstruction and grasping from rgb-d active stereo camera," *arXiv preprint arXiv:2405.05648*, 2024. 2, 5, 6

[37] C. R. A. Chaitanya, A. S. Kaplanyan, C. Schied, M. Salvi, A. Lefohn, D. Nowrouzezahrai, and T. Aila, "Interactive reconstruction of monte carlo image sequences using a recurrent denoising autoencoder," *ACM Transactions on Graphics (TOG)*, vol. 36, no. 4, pp. 1–12, 2017. 4

[38] N. Ahn, B. Kang, and K.-A. Sohn, "Fast, accurate, and lightweight super-resolution with cascading residual network," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 252–268. 4

[39] J. Li, P. Wang, P. Xiong, T. Cai, Z. Yan, L. Yang, J. Liu, H. Fan, and S. Liu, "Practical stereo matching via cascaded recurrent network with adaptive correlation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 16 263–16 272. 4, 7

[40] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2019. 4

[41] G. Xu, X. Wang, X. Ding, and X. Yang, "Iterative geometry encoding volume for stereo matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 21 919–21 928. 5

[42] H. Zhao, H. Zhou, Y. Zhang, J. Chen, Y. Yang, and Y. Zhao, "High-frequency stereo matching network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 1327–1336. 5

[43] X. Wang, G. Xu, H. Jia, and X. Yang, "Selective-stereo: Adaptive frequency information selection for stereo matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 19 701–19 710. 5

[44] B. Wen, M. Trepte, J. Aribido, J. Kautz, O. Gallo, and S. Birchfield, "Foundationstereo: Zero-shot stereo matching," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 5249–5260. 5

[45] H. Jiang, Z. Lou, L. Ding, R. Xu, M. Tan, W. Jiang, and R. Huang, "Defom-stereo: Depth foundation model based stereo matching," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 21 857–21 867. 5

[46] Stereolabs, "Zed 2 stereo camera," https://www.stereolabs.com/en-de/products/zed-2, n.d., accessed: 2025-09-13. 6, 7

[47] H.-S. Fang, M. Gou, C. Wang, and C. Lu, "Robust grasping across diverse sensor qualities: The graspnet-1billion dataset," *The International Journal of Robotics Research*, 2023. 6

[48] X. Guo, J. Lu, C. Zhang, Y. Wang, Y. Duan, T. Yang, Z. Zhu, and L. Chen, "Openstereo: A comprehensive benchmark for stereo matching and strong baseline," 2023. 7

[49] "Middlebury stereo vision page," https://vision.middlebury.edu/stereo/. 7

[50] L. Zhang, K. Bai, G. Huang, Z. Bing, Z. Chen, A. Knoll, and J. Zhang, "Multi-fingered robotic hand grasping in cluttered environments through hand-object contact semantic mapping," *arXiv preprint arXiv:2404.08844*, 2024. 7