2026 IEEE/SICE International Symposium on System Integration (SII)
January 11-14, 2026. Cancún, México

Depth Estimation for Picking Transparent Objects Using a Polarization Camera

Kento Yamada[1], Prasnaht Kumar[1], Yukiyasu Domae[2], Takuya Kiyokawa[1]
Weiwei Wan[1] and Kensuke Harada[1,2]

***Abstract*— For industrial automation, robots have to robustly pick objects with a diverse range of physical properties, such as shape, weight and surface optical properties. To realize such a purpose, this research proposes a method for depth estimation of transparent objects having complex optical properties, such as refraction and reflection from a single viewpoint. While conventional RGB-based or depth-completion approaches struggle to provide reliable predictions of a depth image for such transparent objects, we propose a novel monocular framework that simultaneously estimates the depth and surface normals of transparent objects from a single polarization image. Our method leverages the rich cues provided by polarization and achieves a computationally efficient depth estimation that requires neither analytical models of light reflection nor multi-view setups. To obtain accurate ground-truth labels for a transparent object, the proposed method uses depth and normal maps generated by existing models as pseudo ground-truth, enabling effective learning without manual labels. Experimental results demonstrate that the proposed lightweight framework achieves competitive accuracy in real-world environments.**

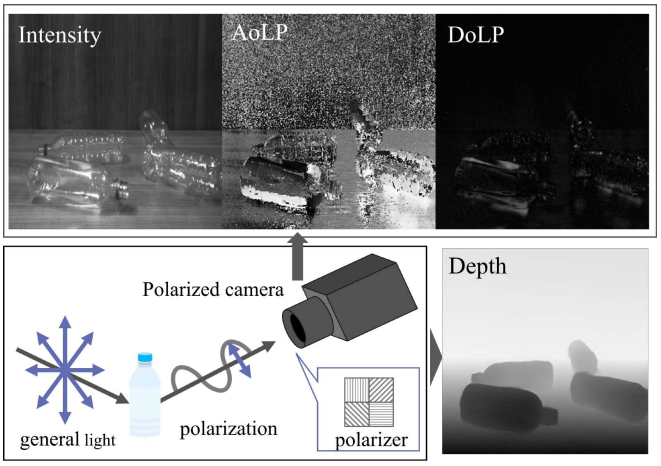

Fig. 1: Depth estimation from a single polarized camera image

I. INTRODUCTION

With the recent advancement of automation in warehouses and factories, industrial robots are increasingly required to recognize and manipulate objects with a wide variety of physical properties, such as shape, weight and surface optical properties. However, recognizing and handling transparent objects like plastic bottles and glassware using vision sensors remains a challenging task. Due to their ability to transmit background information, transparent objects often cause missing or unreliable depth measurements when using conventional depth sensors.

So far, 3D reconstruction methods have been proposed to achieve highly accurate depth and shape estimation by utilizing multiview cameras[1], [2], [3]. However, such approaches require complex and expensive hardware setups and capture the image of the target object from multiple perspectives, making them unsuitable for industrial picking applications. Recently, although monocular RGB sensors have also been used for 3D reconstruction assuming large-scale real-world datasets, the data collection is particularly challenging[4], [5], [6], [7]. On the other hand, this research proposes another 3D reconstruction approach from a monocular camera. Our method utilizes the polarized camera without assuming large-scale real-world datasets. Polarization images offer a promising approach by capturing object surface and material properties that are not obtainable with conventional RGB images or depth sensors. In addition, different from previous methods on 3D reconstruction using polarized camera[8], [9], [10], [11], [12], our method does not need a complex reflection model and simultaneously estimates the depth and surface normals of transparent objects from three polarization images captured by a single polarization camera, as illustrated in Fig. 1. Compared to methods relying solely on RGB images, polarization provides richer cues for transparent objects. We adopt a self-supervised learning strategy for transparent objects, where accurate ground-truth acquisition is inherently difficult. In this setting, we utilize the outputs of existing models [13], [21] as pseudo ground-truth labels for surface normals and depth, respectively. This approach eliminates the need for precise manual annotations and enables large-scale data collection. Importantly, by using different models to generate pseudo labels for surface normals and depth, our method prevents the network from learning the noise patterns specific to a single model. Furthermore, by leveraging the rich information provided by polarization images, the network learns statistical tendencies across the entire dataset and achieves depth estimation accuracy that surpasses the original models.

To summarize, the main contributions of this work are listed as follows:

- We propose a lightweight and computationally efficient depth estimation model for transparent objects that uses only a single polarization image, effectively leveraging

[1]All authors with the superscript 1 are with Osaka University, Osaka, Japan.
[2]Yukiyasu Domae is with the National Institute of Advanced Industrial Science and Technology (AIST), Japan.

polarization cues without requiring complex reflection modeling or multi-view setups, making it suitable for robotic picking tasks.
- We build a framework to incorporate surface normal information as an intermediate output, thereby enhancing depth estimation accuracy.
- We demonstrate that, despite the difficulty of acquiring large-scale real-world polarization datasets, training with pseudo-depth labels enables effective learning of transparent objects without relying on accurate ground-truth annotations.

II. RELATED WORK

A. Depth Estimation

In robotic picking tasks, it is essential to accurately perceive the position and shape of target objects. Although multi-view methods, such as NeRF [1], [2], [3], enable highly accurate 3D reconstruction from multiple viewpoints, these methods require specialized hardware and significant computational cost. In contrast, depth estimation from monocular camera has also been proposed[13], [7], [21], [4]. A major challenge in depth estimation lies in handling transparent objects. ClearGrasp [13], NDDepth [7] and Liu et al.[15] proposed methods for depth estimation of transparent objects from RGB images. However, the performance of the depth estimation of transparent objects is limited just by using the RGB information. Recently, the transformer-based models[4], [16], [17], [18], [19] have also demonstrated the ability to cope with reflections and refractions. However, this method needs a large amount of training data which are difficult to obtain.

B. Shape Estimation Using Polarization Images

Polarization images provide valuable cues for inferring object geometry. Classical Shape-from-Polarization (SfP) approaches combine polarization information with reflection models to estimate surface normals, including those of transparent objects [9], [8]. More recent studies extend these methods by incorporating attention mechanisms and photometric fusion under natural illumination to handle more complex scenes [10], [24]. Polarization cues have also proven effective for segmentation tasks, particularly for detecting transparent and specular objects [25], [26]. In the field of depth estimation, several approaches have been explored using polarization images, such as stereo-based depth estimation [11], depth completion [14], and structured polarized illumination [12]. However, these methods rely on multiple cameras or specialized hardware and are limited by the complexity of physical models, multi-view dependencies, and high computational and hardware costs. As far as we know, this is the first trial of depth estimation of transparent objects from a single viewpoint where such a method can be well applied to the robotic picking task.

Obtaining high-quality depth labels is often costly and labor-intensive. To mitigate this issue, recent studies have investigated the use of pseudo labels [20], [21] as well as self-supervised and weakly supervised learning methods [6], [17], [22], [23], which enable depth estimation models to achieve high accuracy without relying on manual ground-truth labels. In this research, we also provide an efficient method to use the pseudo labels for transparent objects.

III. METHOD

The proposed framework consists of two stages. In the first stage, we construct a deep learning model that takes three polarization images captured with a monocular polarization camera as input. This model estimates the surface-normal, distance, depth, and uncertainty maps as intermediate outputs, where the distance map denotes the distance from the camera center to each planar surface within the scene. In the second stage, we subsequently predict the final depth map from the intermediate outputs of the first stage. The first stage is founded on the assumption that real-world scenes can be represented as a collection of planes. Rather than directly predicting depth, we estimate surface-normal and distance maps as intermediate representations, which provide stable, piecewise-constant information for each plane. From these intermediate outputs, the final depth image is estimated in the second stage. This formulation enables depth estimation that explicitly exploits the underlying planar structure of the scene. This is considered to be more accurate than the intermediate depth maps generated in the first stage.

A. Polarization Image Formation

Polarization is defined as the state in which the oscillation direction of a light wave is aligned along a specific orientation. Because the surface reflectance of an object depends on both the angle of incidence and the polarization direction, the reflected light becomes partially polarized. This phenomenon provides complementary information that cannot be obtained from RGB or depth images alone. In this study, a Baumer polarized camera[33] is used to capture images at four polarization angles (I_{0°, I_{45°, I_{90°, and I_{135°). From these images, the Stokes parameters S_0, S_1, and S_2 are computed as

$$S_0 = I_{0^\circ} + I_{90^\circ}, \tag{1}$$

$$S_1 = I_{0^\circ} - I_{90^\circ}, \tag{2}$$

$$S_2 = I_{45^\circ} - I_{135^\circ}. \tag{3}$$

Using these parameters, the Intensity (I), Angle of Linear Polarization (AoLP) and Degree of Linear Polarization (DoLP) shown in Fig. 1 are calculated as

$$\mathrm{I} = \frac{I_{0^\circ} + I_{45^\circ} + I_{90^\circ} + I_{135^\circ}}{4}, \tag{4}$$

$$\mathrm{DoLP} = \frac{\sqrt{S_1^2 + S_2^2}}{S_0}, \tag{5}$$

$$\mathrm{AoLP} = \frac{1}{2}\tan^{-1}\left(\frac{S_2}{S_1}\right). \tag{6}$$

These three polarized images are used as input to the proposed network.

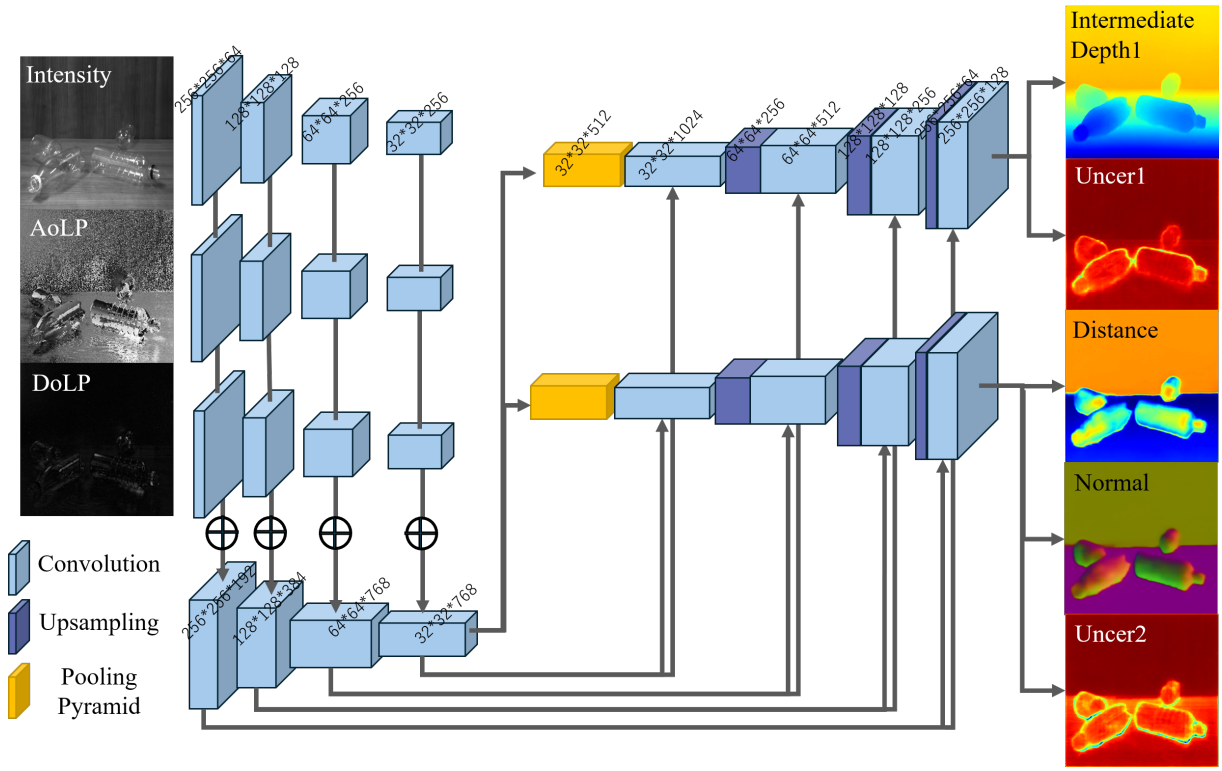

Fig. 2: Network structure to generate intermediate outputs.

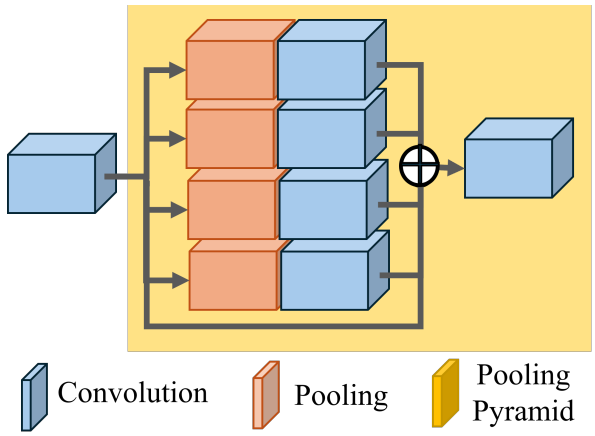

Fig. 3: The structure of pooling pyramid.

B. Network Architecture

The proposed method for depth estimation is composed of two stages: generation of the intermediate outputs and generation of the final depth image. Below, we describe each stage in more detail.

Generation of the intermediate outputs: The network structure for generating intermediate outputs is illustrated in Fig. 2. In this stage, the polarization image is processed through a CNN to produce intermediate outputs, including depth, surface normals, distance, and uncertainty maps. The model adopts an encoder structure inspired by RGBD2Normal[27] and a decoder structure based on ND-Depth [7]. Since three input branches are required, and model lightweighting is essential, we chose the encoder from RGBD2Normal due to its design emphasis on efficiency and compactness. In addition, the encoder should be designed to fuse features within the decoder, which results in having two decoders per single encoder. Consequently, this leads to a complex model architecture with three encoder branches and six decoder branches. In contrast, our approach differs from RGBD2Normal by performing feature aggregation at the encoder stage, resulting in a simpler model configuration with three encoder branches and two decoder branches. The encoder adopts an FCN architecture based on VGG-16. To reduce the model size, the final block is removed, and the number of channels in the conv4 block is reduced from 512 to 256. Features extracted from the three input branches are summed at each stage, resulting in encoder feature dimensions of [192, 384, 768, 768]. Skip connections are introduced to preserve local features. The features obtained from the encoder are processed, as illustrated in Fig. 3, by performing average pooling at four different scales (1, 2, 3, and 6) to capture both local and global contextual information, followed by convolution and upsampling operations. Subsequently, the processed features are fed into two separate decoders: one for intermediate depth map estimation and the other for surface normal and distance map estimation. Each decoder has an output head at its final layer to generate the corresponding maps.

Generation of the final depth map: The process of generating the final depth map from the intermediate outputs

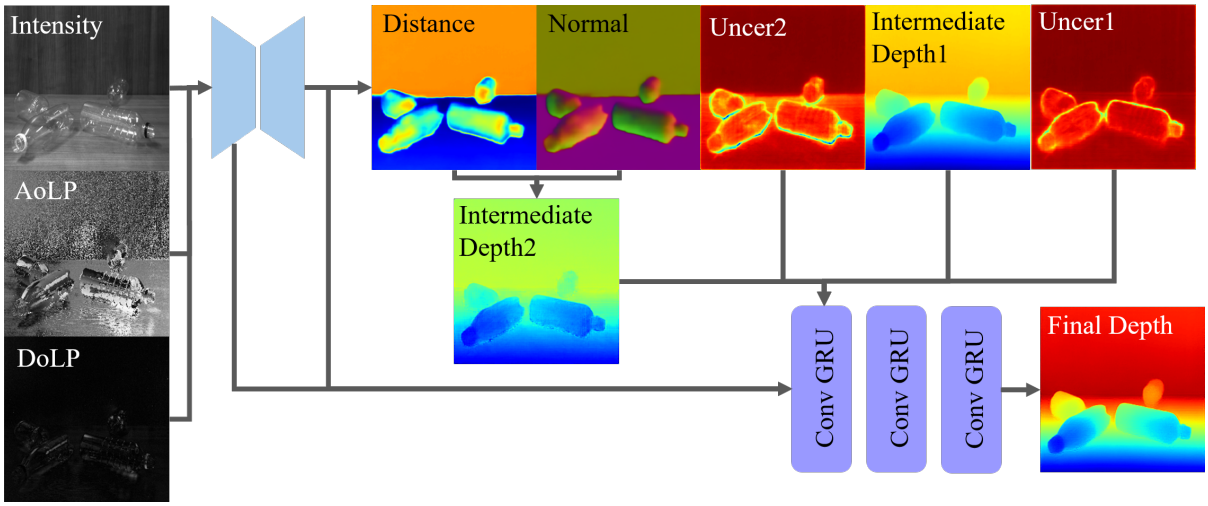

Fig. 4: Network structure to generate the final depth map.

of the first stage is illustrated in Fig. 4. This procedure is identical to that used in NDDepth. We first further generate the intermediate depth image (Intermediate depth 2 in Fig. 4) derived from the normal and distance maps by using the following equation:

$$\text{Depth}(p) = \frac{D(p)}{N(p) \cdot K^{-1} \tilde{p}}, \tag{7}$$

where $N(p)$, $D(p)$, K, p and $\tilde{p}$ denote the surface-normal map, distance map, pixel coordinate, and camera intrinsic matrix, and pixel coordinates in homogeneous form, respectively. This intermediate depth map is generated by assuming the planar structure of the scene. On the other hand, the intermediate depth map (Intermediate depth 1 in Fig. 4) generated in stage 1 does not assume the planar structure of the scene, where it will be more reliable in areas with high curvature. These inputs are weighted according to their uncertainty maps. Then, the combined features obtained through an encoder and decoder are further processed by a three-stage GRU module. The GRU progressively integrates and refines this information, gradually improving the initial depth map and producing the final depth map.

C. Training Dataset and Pseudo Ground-Truth

Polarization is highly sensitive and varies substantially depending on the geometry and material properties of surrounding objects. Consequently, polarization datasets generated in simulated environments exhibit significant domain gaps compared to those captured in real-world scenes, rendering direct use of such simulated data for training purposes challenging. Moreover, acquiring ground-truth depth data for real-world scenes containing transparent objects remains difficult due to missing regions in depth images and the necessity of precise calibration between depth and polarization cameras. While some approaches replace transparent objects with colored, semi-transparent proxies solely during depth acquisition, this process is entirely manual, incurring considerable time and labor costs, thereby hindering large-scale data collection.

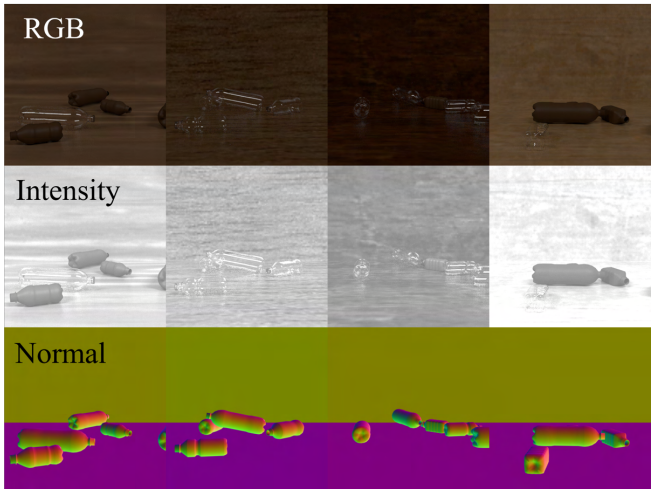

Fig. 5: Synthetic datasets

To address these challenges, this study leverages a real-world dataset wherein pseudo ground-truth is synthesized via existing models, comprising estimated surface normals and depth maps from scenes with transparent objects, alongside depth maps from corresponding scenes in which transparent objects are substituted with opaque counterparts. This approach enables training with labels corresponding to real images, eliminating the need for precise manual annotations and allowing efficient collection of large-scale training data. By employing separate models to generate pseudo labels for surface normals and depth, the network is prevented from overfitting to noise patterns specific to a single model. Moreover, by incorporating the rich information provided by polarization images, the network can capture statistical tendencies across the entire dataset, achieving depth estimation accuracy that surpasses the original models. In the training process, only depth maps and surface-normal maps are used as ground-truth labels, while the distance maps are implicitly generated within the model by computing

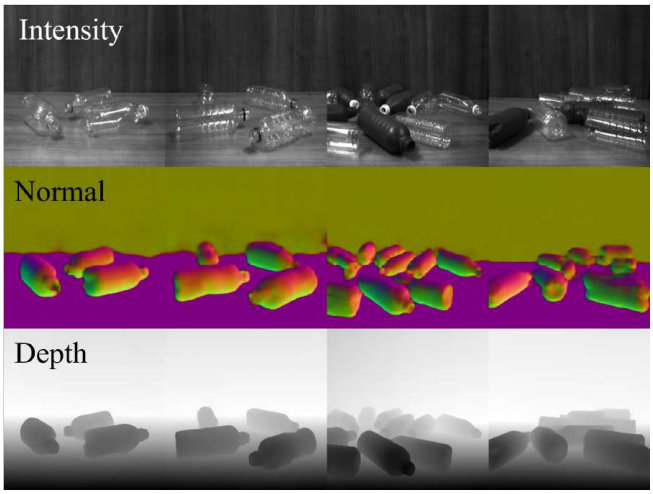

Fig. 6: Estimated pseudo dataset

them from the ground-truth depth and surface-normal labels.

Pseudo normal labels:

The pseudo-normal labels are generated from a simulation environment. We utilized a synthetic dataset in which scenes generated in Unity [31] were rendered in Blender [30] using Mitsuba3 [32] as the rendering engine. Here, we observed that, among AoLP, DoLP and Intensity images, the domain gap of the Intensity image which represents the average luminance across four polarization directions is the smallest. We decided to use the deep learning model ClearGrasp [13] originally designed to predict surface normal maps from single RGB images to estimate the pseudo-normal labels from the intensity image. We fine-tuned the pretrained checkpoint of the ClearGrasp model using Intensity images as input. Each scene contained three to five randomly placed transparent and opaque plastic bottles. Examples of the synthetic dataset and normal maps predicted by the fine-tuned model are shown in Fig. 5 and Fig. 6.

Pseudo depth labels: For depth label estimation, we use DepthAnythingV2 [21] additionally trained with spray-painted objects, where DepthAnythingV2 was originally designed to predict depth maps from monocular RGB images. Although originally intended for RGB inputs, we confirmed that the model could also produce valid depth maps when provided with Intensity images. Therefore, we used the pretrained checkpoint without modification to estimate depth maps directly from Intensity images. However, as shown in Fig. 7, the depth maps predicted by DepthAnythingV2 still exhibit room for improvement, with issues such as missing regions and the inability to represent the bottle cap and body as a continuous surface.

To reduce estimation errors in transparent regions, we prepared two types of pseudo-depth labels: (1) those directly estimated from scenes containing transparent objects, and (2) in addition, as illustrated in Fig. 7, those obtained from corresponding scenes in which the transparent objects were replaced with identically shaped opaque counterparts created by spray-painting. For the latter, in order to maintain the same object pose during replacement, we 3D-printed custom fixtures to ensure consistent placement of the opaque objects, thereby enabling more accurate acquisition of pseudo labels.

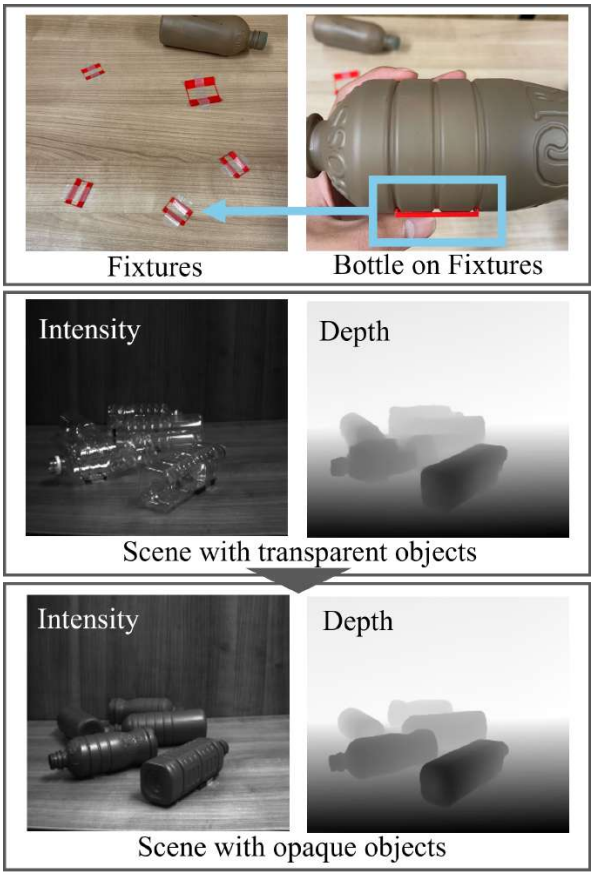

Fig. 7: Accurate pseudo ground-truth depth generation

IV. Experiment

We evaluate the ability of the proposed method to estimate the geometry of transparent objects on a real-world dataset. We also performed the picking experiment of transparent objects where the depth is estimated by our proposed method.

Training: This model is executed on a T4 GPU in Google Colab. For gradient computation, the Adam optimizer [28] is employed with a batch size of 8. The learning rate follows a polynomial decay schedule, decreasing from an initial value of 2×10^{-5} to 2×10^{-6}. Training is conducted for 100 epochs. We used vacant pet bottles as objects having transparency. Note that the polarization images and RGB images are acquired using different cameras, and therefore their fields of view do not perfectly match.

A. Depth Estimation Experiments

We quantitatively evaluate the depth estimation results of the proposed method. First, we describe the metrics and baseline methods we used, followed by the presentation of the estimation results.

Metrics: We follow previous studies[5], [13] and employ the following evaluation metrics:

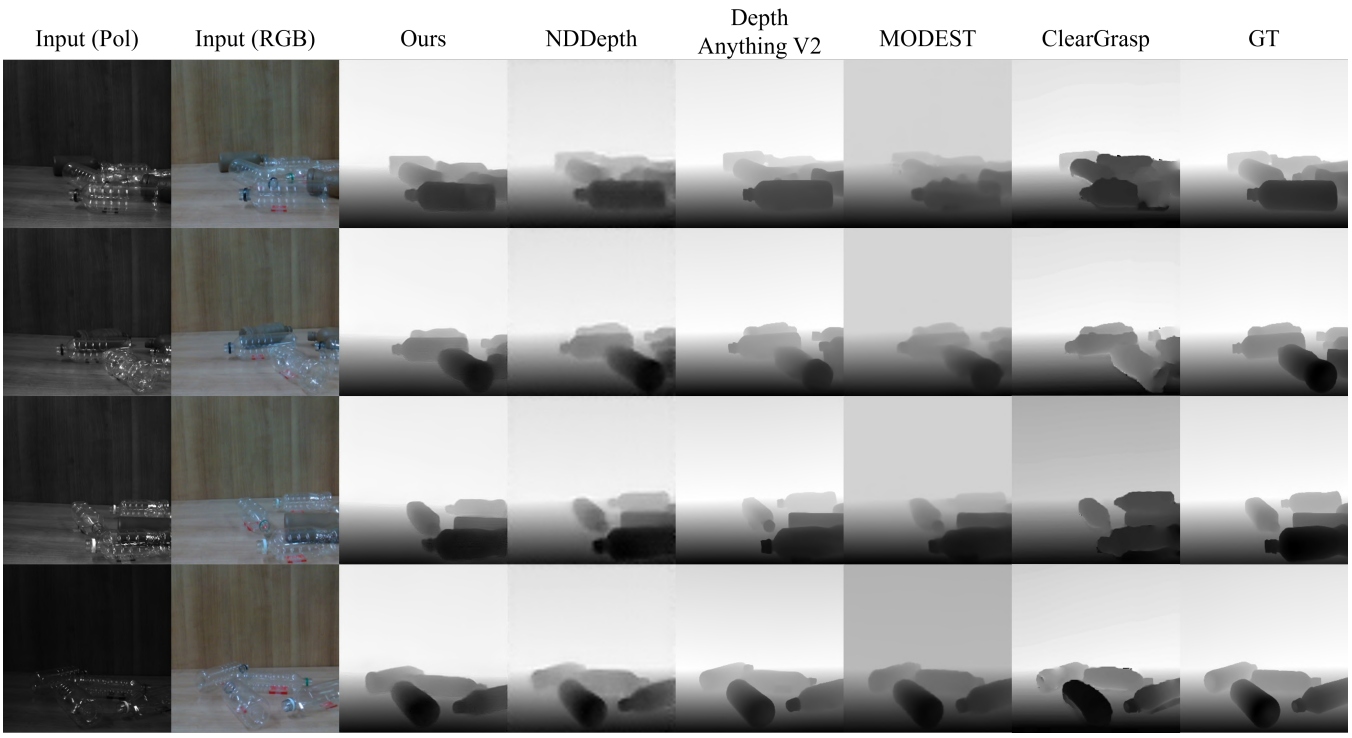

Fig. 8: Depth estimation results

TABLE I: Quantitative comparison of depth estimation methods evaluated on the transparent regions of the image

Method	MAE↓	RMSE↓	REL↓	$\delta_{1.05}\uparrow$	$\delta_{1.10}\uparrow$	$\delta_{1.25}\uparrow$
Ours	**0.028**	**0.033**	**0.406**	18.18	**38.06**	**71.23**
[21]	0.053	0.060	0.820	13.50	25.03	49.24
[7]	0.055	0.061	0.805	9.38	19.59	45.13
[15]	0.047	0.056	0.642	12.70	24.13	47.05
[13]	0.080	0.116	4.668	**29.35**	35.62	47.37

TABLE II: Quantitative comparison of depth estimation methods over the entire image.

Method	MAE↓	RMSE↓	REL↓	$\delta_{1.05}\uparrow$	$\delta_{1.10}\uparrow$	$\delta_{1.25}\uparrow$
Ours	0.019	0.024	0.114	58.15	80.34	**94.01**
[21]	**0.013**	**0.022**	**0.112**	**78.00**	**85.36**	91.90
[7]	0.030	0.035	0.188	38.19	72.80	89.23
[15]	0.053	0.066	0.237	22.42	37.64	75.47

- **RMSE**: Root Mean Squared Error between predicted and ground-truth depth values (in meters).
- **MAE**: Mean Absolute Error between predicted and ground-truth depth values (in meters).
- **REL**: Mean Absolute Relative Error, calculated as the average of $\frac{|d-d^*|}{d^*}$ over all valid pixels.
- δ_x: The percentage of pixels for which $\max\left(\frac{d}{d^*}, \frac{d^*}{d}\right) < x$, where $x \in \{1.05, 1.10, 1.25\}$.

Comparative methods: As comparative methods, we use the monocular depth estimation approaches ClearGrasp[13], NDDepth[7], Modest[15], and DepthAnythingV2[21]. We trained each model using 3,600 sets of training data. Only DepthAnythingV2 utilizes a pre-trained checkpoint learned from an existing large-scale dataset.

Result: We performed depth estimation on 58 different scenes. The estimation results are shown in Fig. 8. Our method demonstrates particularly high estimation accuracy in regions where transparent objects overlap and along their boundaries. Quantitative evaluations were conducted under two settings: (i) restricted to the regions corresponding to transparent objects, and (ii) over the entire scene, including both transparent and opaque regions. The results are summarized in Tables I and II. Since ClearGrasp [13] is designed as a depth-completion method, its evaluation is limited to transparent regions only. For the evaluation restricted to transparent regions (Table I), the proposed method achieves substantially lower MAE, RMSE and REL values and higher scores on the δ metrics, indicating strong capability in handling challenging transparent surfaces. Although ClearGrasp [13] attains the highest performance in terms of the $\delta_{1.05}\uparrow$ metric, its MAE, RMSE and REL errors are more than three times larger than those of the proposed method, suggesting that its depth estimations are less accurate. In the evaluation performed on the entire images (Table II), DepthAnythingV2 outperforms all competing methods on all metrics except $\delta_{1.05}\uparrow$. Nevertheless, the proposed method consistently surpasses the other approaches across all metrics, while employing only approximately one-third of the parameters of DepthAnythingV2 [21]. These results demonstrate that the proposed method can achieve efficient and accurate reconstruction of object geometry with

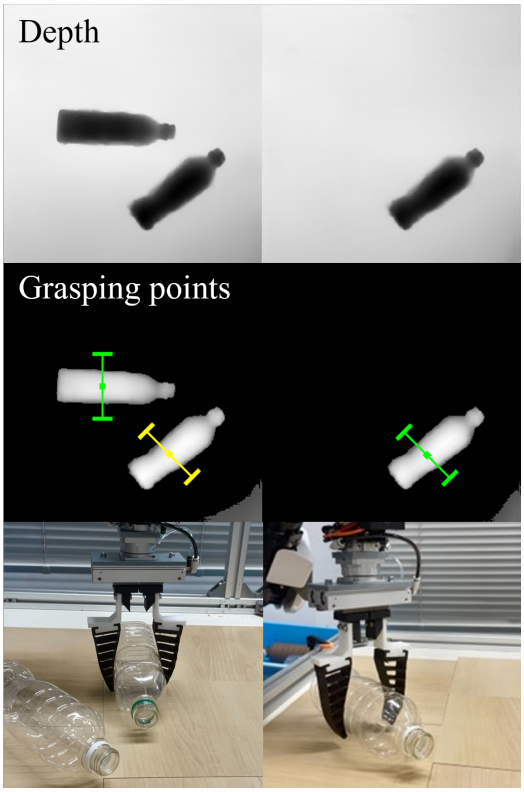

Fig. 9: Picking experiment of transparent objects.

a significantly more compact model.

B. Picking experiment

We apply the proposed method to a robotic picking system to demonstrate its effectiveness in performing tasks involving transparent objects. First, we describe the experimental setup, followed by the presentation of the experimental results.

Experimental Setup: We applied the proposed algorithm to a robotic picking task and conducted experiments involving transparent objects. The robot platform used was Nextage Open from Kawada Robotics Inc. In the experimental setup, two or three transparent objects were placed on a table within the robot's workspace, and polarization images were captured using a Baumer camera. The objective of the robot is to grasp objects on the table and place them into a box positioned beside the robot, following the proposed grasping algorithm. The end-effector is equipped with a Fin Ray Finger, and the grasping position estimation system is implemented using Fast-Graspability [29].

Metrics: For each method, we conducted 30 trials and recorded the number of successful attempts (#Succ) as well as two types of failures: #F-Cont and #F-GP. #F-Cont refers to failures in which the gripper made unintended contact with the table exceeding 1 cm, despite following a grasp point determined from the estimated depth image. #F-GP corresponds to failures caused by incorrect specification of the grasp point. The success rate (Succ%) is calculated as the number of successful attempts divided by the total number of trials: Succ% = #Success / #Attempts.

TABLE III: Picking experiment results

Meth	Obj	#Attempts	#Succ	#F-Cont	#F-GP	Succ%
Ours	T	30	28	0	2	96.0
[21]	T	30	22	5	3	73.3
[7]	T	30	19	9	2	63.3
[15]	T	30	24	5	1	80.0
Ours	O	30	29	0	1	96.7

Comparative methods: We use DepthAnythingV2, ND-Depth, and Modest as baseline methods for comparison. To further evaluate the robustness of the proposed method, additional grasping experiments were conducted on opaque objects. We used only 200 sets of pseudo labels directly estimated from scenes containing transparent objects as training data.

Result: An overview of the experimental setup is shown in Fig. 9. A depth image is estimated from the images captured by a camera mounted perpendicularly above the table, and subsequently passed to Fast-Graspability. Multiple candidate grasp points are generated, and the one with the highest predicted success rate is selected for the pick-and-place operation. The selected grasp point is shown in green, while the remaining candidates are visualized in yellow. The experimental results are summarized in Table III. The object type is denoted as "T" for transparent objects and "O" for opaque objects. In the experiments on transparent objects, the proposed method achieved the highest success rate among all evaluated approaches. Notably, unlike other methods including DepthAnythingV2, no failures due to gripper contact with the table were observed. This result suggests that, despite being trained with pseudo datasets, our method is capable of accurately estimating the depth of transparent objects with the correct scale. Moreover, when compared with the results of the experiments on opaque objects, the proposed method achieves a comparable level of precision in executing the picking task, as indicated by both the task success rate and the number of failures attributed to incorrect grasp point specification.

V. CONCLUSIONS

In this paper, we propose a lightweight depth estimation model for transparent objects that relies solely on a single polarization image, providing a practical solution for robotic picking tasks without the need for complex hardware. Additionally, our training method using pseudo depth labels addresses the annotation scarcity in polarization images, enabling scalable learning for transparent object recognition. In the future, we aim to extend the applicability of our approach beyond PET bottles to various other objects, including more complex scenes, in order to realize a more general and robust method for diverse picking tasks.

ACKNOWLEDGEMENT

This work was partially supported by the JST-SICORP project.

REFERENCES

[1] B. P. Duisterhof, Y. Mao, S. H. Teng, and J. Ichnowski, "Residual-NeRF: Learning Residual NeRFs for Transparent Object Manipulation," in *Proc. IEEE International Conference on Robotics and Automation (ICRA)*, Yokohama, Japan, May 2024, pp. 13918–13924.

[2] J. Ichnowski, Y. Avigal, J. Kerr, and K. Goldberg, "Dex-NeRF: Using a Neural Radiance Field to Grasp Transparent Objects," in *Proc. Conference on Robot Learning (CoRL)*, Cambridge, MA, USA, Nov. 2020.

[3] A. Ummadisingu, J. Choi, K. Yamane, S. Masuda, N. Fukaya, and K. Takahashi, "SAID-NeRF: Segmentation-AIDed NeRF for Depth Completion of Transparent Objects," in *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Abu Dhabi, United Arab Emirates, Oct. 2024, pp. 7535–7542.

[4] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler, and V. Koltun, "Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-Shot Cross-Dataset Transfer," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 3, pp. 1623–1637, Mar. 2022.

[5] 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, Jul. 2022.

[6] M. Blanchon, D. Sidibé, O. Morel, R. Seulin, D. Braun, and F. Meriaudeau, "P2D: A Self-Supervised Method for Depth Estimation from Polarimetry," in *Proc. 25th International Conference on Pattern Recognition (ICPR)*, Milan, Italy, Jan. 2021, pp. 7357–7364.

[7] S. Shao, Z. Pei, W. Chen, P. C. Y. Chen, and Z. Li, "NDDepth: Normal-Distance Assisted Monocular Depth Estimation and Completion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 12, pp. 8883–8899, Dec. 2024.

[8] M. Shao, C. Xia, Z. Yang, J. Huang, and X. Wang, "Transparent Shape from a Single View Polarization Image," in *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, Paris, France, Oct. 2023, pp. 9243–9252.

[9] Y. Ba, A. Gilbert, F. Wang, J. Yang, R. Chen, Y. Wang, L. Yan, B. Shi, and A. Kadambi, "Deep Shape from Polarization," in *European Conference on Computer Vision (ECCV)*, Glasgow, UK, Aug. 2020, pp. 554–571, Springer.

[10] C. Lei, C. Qi, J. Xie, N. Fan, V. Koltun, and Q. Chen, "Shape from Polarization for Complex Scenes in the Wild," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, LA, USA, Jun. 2022, pp. 12622–12631.

[11] C. Tian, W. Pan, Z. Wang, M. Mao, G. Zhang, H. Bao, P. Tan, and Z. Cui, "DPS-Net: Deep Polarimetric Stereo Depth Estimation," in *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, Paris, France, Oct. 2023, pp. 3546–3556.

[12] T. Ichikawa, S. Nobuhara, and K. Nishino, "SPIDeRS: Structured Polarization for Invisible Depth and Reflectance Sensing," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, Jun. 2024, pp. 25077–25085.

[13] 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 *Proc. IEEE International Conference on Robotics and Automation (ICRA)*, Paris, France, May 2020, pp. 3634–3642.

[14] K. Ikemura, Y. Huang, F. Heide, Z. Zhang, Q. Chen, and C. Lei, "Robust Depth Enhancement via Polarization Prompt Fusion Tuning," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, Jun. 2024, pp. 20710–20720.

[15] 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 *Proc. IEEE Int. Conf. Robotics and Automation (ICRA)*, Yokohama, Japan, 2025, to be published.

[16] C. Godard, O. M. Aodha, and G. J. Brostow, "Unsupervised Monocular Depth Estimation with Left-Right Consistency," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 6602–6611.

[17] C. Godard, O. M. Aodha, M. Firman, and G. Brostow, "Digging Into Self-Supervised Monocular Depth Estimation," in *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, South Korea, Oct. 2019, pp. 3827–3837.

[18] B. Ke, A. Obukhov, S. Huang, N. Metzger, R. C. Daudt, and K. Schindler, "Repurposing Diffusion-Based Image Generators for Monocular Depth Estimation," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, Jun. 2024, pp. 9492–9502.

[19] K. Chen, S. Wang, B. Xia, D. Li, Z. Kan, and B. Li, "TODE-Trans: Transparent Object Depth Estimation with Transformer," in *Proc. IEEE International Conference on Robotics and Automation (ICRA)*, London, UK, May 2023, pp. 4880–4886.

[20] A. Costanzino, P. R. Zama, M. Poggi, F. Tosi, S. Mattoccia, and L. Di Stefano, "Learning Depth Estimation for Transparent and Mirror Surfaces," in *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, Paris, France, Oct. 2023, pp. 9210–9221.

[21] Z. Hu, H. Wang, X. Rong, D. Zhang, Z. Xu, and F. Guan, "Integrating Depth-Anything-V2 Depth Estimation and Sobel Operator Matrix for UAV Landing-site Detection," in *Proc. International Conference on Cyberworlds (CW)*, Tokyo, Japan, Sep. 2024, pp. 342–343.

[22] V. Guizilini, R. Ambrus, S. Pillai, A. Raventos, and A. Gaidon, "3D Packing for Self-Supervised Monocular Depth Estimation," arXiv preprint arXiv:1905.02693, 2020.

[23] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, "Unsupervised Learning of Depth and Ego-Motion from Video," in *Proc. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, 2017, pp. 6612–6619.

[24] Y. Ding, Y. Ji, and J. Ye, "Polar-Photometric Stereo Under Natural Illumination," in *Proc. International Conference on 3D Vision (3DV)*, Athens, Greece, Oct. 2022, pp. 690–699.

[25] H. Mei, B. Dong, W. Dong, J. Yang, S.-H. Baek, F. Heide, P. Peers, X. Wei, and X. Yang, "Glass Segmentation using Intensity and Spectral Polarization Cues," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, LA, USA, Jun. 2022, pp. 12612–12621.

[26] A. Kalra, V. Taamazyan, S. K. Rao, K. Venkataraman, R. Raskar, and A. Kadambi, "Deep Polarization Cues for Transparent Object Segmentation," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, Jun. 2020, pp. 8599–8608.

[27] J. Zeng, Y. Tong, Y. Huang, Q. Yan, W. Sun, J. Chen, and Y. Wang, "Deep Surface Normal Estimation With Hierarchical RGB-D Fusion," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 6146–6155.

[28] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *CoRR*, vol. abs/1412.6980, 2014.

[29] Y. Domae, H. Okuda, Y. Taguchi, K. Sumi, and T. Hirai, "Fast graspability evaluation on single depth maps for bin picking with general grippers", in *Proc. IEEE International Conference on Robotics and Automation (ICRA)*, 2014, pp. 1997–2004.

[30] Blender Online Community, *Blender – a 3D modelling and rendering package*, Blender Foundation. [Online]. Available: `https://www.blender.org/`

[31] Unity Technologies, *Unity Real-Time Development Platform*. [Online]. Available: `https://unity.com/`

[32] W. Jakob et al., *Mitsuba 3 Renderer*, 2023. [Online]. Available: `https://github.com/mitsuba-renderer/mitsuba3`

[33] Baumer, *VCXG-50MP 5MP Polarization Camera*. [Online]. Available: `https://www.argocorp.com/cam/Gige/VCXG_Baumer/VCXG-50MP.html`. [Accessed: Aug. 11, 2025]