

Monocular Depth Estimation and Segmentation for Transparent Object with Iterative Semantic and Geometric Fusion

Jiangyuan Liu^{1,2}, Hongxuan Ma^{1,2}, Yuxin Guo^{1,2}, Yuhao Zhao^{1,2}, Chi Zhang³, Wei Sui⁴, Wei Zou^{1,2†}

Abstract—Transparent object perception is indispensable for numerous robotic tasks. However, accurately segmenting and estimating the depth of transparent objects remain challenging due to complex optical properties. Existing methods primarily delve into only one task using extra inputs or specialized sensors, neglecting the valuable interactions among tasks and the subsequent refinement process, leading to suboptimal and blurry predictions. To address these issues, we propose a monocular framework, which is the first to excel in both segmentation and depth estimation of transparent objects, with only a single-image input. Specifically, we devise a novel semantic and geometric fusion module, effectively integrating the multi-scale information between tasks. In addition, drawing inspiration from human perception of objects, we further incorporate an iterative strategy, which progressively refines initial features for clearer results. Experiments on two challenging synthetic and real-world datasets demonstrate that our model surpasses state-of-the-art monocular, stereo, and multi-view methods by a large margin of about 38.8%-46.2% with only a single RGB input. Codes and models are publicly available at <https://github.com/L-J-Yuan/MODEST>.

I. INTRODUCTION

Transparent objects such as bottles, flasks, and windows are ubiquitous in various domains, like laboratories, industries, or daily life. For robots in these scenarios, accurately detecting and estimating the depth of transparent objects are usually prerequisites for subsequent manipulation and navigation tasks [1]. However, transparent objects often lack clear texture and blend with the background in most RGB images, due to their complex refraction and reflection characteristics [2]. Additionally, commercial depth cameras also struggle to perceive such objects [3], thus producing incomplete and noisy depth maps. These failures of conventional sensors hinder the development of downstream tasks like grasping.

Confronting these issues, previous researches independently focus on either segmentation or depth estimation of transparent objects, using supplementary modalities, as shown in Fig. 1. For example, some works resort to specialized sensors, such as polarized camera [4], RGB-Thermal camera [5], etc., which are typically expensive and difficult to obtain. Other methods utilize multi-view RGB images [6]–[8] or additional depth maps [9]–[11] as inputs, leading to substantial time overhead and suboptimal performance

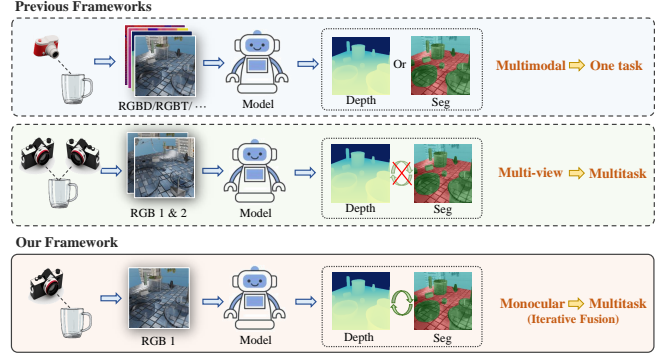


Fig. 1. Previous frameworks rely either on multi-view inputs or additional modalities (e.g., depth maps, thermal images) to make predictions. Differently, we propose the first monocular framework that utilizes iterative cross-task fusion to improve both depth and segmentation performance.

due to incomplete and noisy depth maps of transparent objects. More recently, SimNet [12] and MVTrans [13] adopt multi-task frameworks for transparent object perception in stereo and multi-view settings, respectively. However, they overlook the beneficial information and interactions across multiple tasks, resulting in notably inferior and unbalanced performance of both segmentation and depth estimation.

To address these issues, we analyze and identify two key breakthroughs. **(a) Integrating semantic and geometric interactions into complementary tasks can fully exploit the useful mutual information.** As an ill-posed problem, monocular depth estimation is particularly challenging for transparent objects. Fortunately, semantic segmentation is informative for depth estimation by offering semantic and contextual clues [14]. Similarly, depth estimation could also provide valuable multi-scale geometric information for segmentation, such as boundaries, surfaces, and shapes to assist in determining semantic categories [15]. **(b) Iterating multi-scale fusion continuously can refine the initial fusion results.** When humans observe inconspicuous objects, we tend to notice the overall outline of the object first, then the local details [16]. Inspired by this, we believe that updating features in a coarse-to-fine fashion facilitates transparent object perception.

Based on our analysis, we for the first time propose a monocular framework to concurrently predict precise segmentation and depth for transparent objects. Different from previous works, we take the simplest and most efficient form using only a single RGB input, as shown in Fig. 1. Specifically, to fully exploit the complementary information across tasks, we design a novel semantic and geometric

¹School of Artificial Intelligence, University of Chinese Academy of Sciences

²State Key Laboratory of Multimodal Artificial Intelligence Systems (MAIS), Institute of Automation of Chinese Academy of Sciences

³School of Information Science and Technology, Shijiazhuang Tiedao University

⁴D-Robotics

[†]Corresponding to wei.zou@ia.ac.cn

fusion module that adaptively interacts with the features of both tasks, allowing the model to effectively enhance the predictions, especially for depth estimation. Moreover, to obtain more fine-grained and accurate predictions, we propose an iterative strategy to repeatedly update the initial features through a shared decoder, thereby further improving the performance of both tasks. Extensive experiments on both synthetic and real-world datasets show that, our model is superior to general multi-task methods, and outperforms state-of-the-art stereo and multi-view methods significantly in both depth and segmentation for transparent objects.

In summary, our main contributions are as follows:

- To the best of our knowledge, we propose the first end-to-end monocular framework excelling in predicting both depth and segmentation for transparent objects.
- The key advantages of our approach lie in the semantic and geometric fusion module and an innovative iterative strategy, which better leverage the complementary information between the two tasks, significantly improving transparent object perception.
- Experimental results demonstrate that our model outperforms state-of-the-art monocular and even multi-view methods by a large margin quantitatively and qualitatively, on both synthetic and real datasets.

II. RELATED WORK

A. Transparent Object Segmentation

Accurate detection or segmentation is usually the first step in perceiving and manipulating untextured transparent objects. On the one hand, many existing works utilize specific visual cues to segment transparent objects. For instance, TransLab [17] and EBLNet [18] demonstrated the effectiveness of boundaries for locating transparent objects. GDNet [19] and RFENet [20] proposed novel feature fusion modules to enhance performance by better utilizing contextual and reciprocal features, respectively. On the other hand, some methods obtain additional information gains by means of different input modalities. PGSNet [4] employed a polarized camera to extract optical cues beneficial for segmentation. In [5], thermal images were combined by a multi-modal fusion module to assist in detecting glass surfaces. Differently, our method only takes a single RGB image as input, without relying on additional modalities.

B. Transparent Object Depth Estimation

The techniques for depth estimation of transparent objects can be roughly classified into depth completion and NeRF-based methods. ClearGrasp [3] pioneered the use of RGB-D input for transparent object depth completion. Successive improvements have come from LIDF-Refine [21] and TransparentNet [22] by lifting depth maps to point clouds and performing completion. A more recent work TODE [11] leveraged swin transformer [23] to better capture the global information. Following recent advancements in NeRF [24], DexNeRF [6] and EvoNeRF [7] employed implicit functions to represent transparent objects, though the optimization processes were time-consuming. GraspNeRF [8]

and ResidualNeRF [25] later sped up inference by utilizing the generalizable NeRF and decoupling the background, respectively. Most methods predict depth only once, while we take an iterative way for further refinement.

C. Multi-task Predictions for Transparent Objects

Multi-task dense predictions aim to learn multiple tasks jointly in a unified framework [26]–[28]. ClearGrasp [3] adopted edges, masks and surface normals as intermediate representations for optimizing depth. SimNet explored a multi-task framework based on stereo input to support transparent object manipulation, while recently MVTrans [13] extended it by introducing multiple views. However, none of the above methods for transparent objects leveraged inter-task interactions. In contrast, we propose a fusion module to fully exploit the complementary information between different tasks.

III. METHOD

A. Problem Statement and Method Overview

Given a single RGB image $I \in R^{3 \times H \times W}$, where H is the height and W is the width of the image, the objective is to obtain an accurate segmentation mask $S \in R^{N \times H \times W}$ and a depth map $D \in R^{H \times W}$ for transparent objects, where N is the number of semantic categories. Our model learns a function f that maps the input to two outputs, defined as $(S, D) = f(I)$.

As depicted in Fig. 2, the overall architecture is composed of a transformer-based encoder, a reassemble module and an iterative fusion decoder. In the encoder, the input RGB image is first processed and passed through multiple transformer blocks to extract features as vision tokens. Then we assemble tokens from different layers into multi-scale feature maps, which form two feature pyramids for depth and segmentation, respectively. In the decoder, the two branches are merged together through our semantic and geometric fusion module. This multi-scale fusion and decoding process is iteratively refined through gated units several times to obtain the final depth and segmentation predictions.

B. Transformer Encoder

Existing methods dealing with transparent objects mostly utilize CNN as feature extractors [3], [13]. However, we argue that compared with traditional convolution operators, attention mechanisms provide global contextual representation, which has proven to be especially effective for transparent objects [11]. Thus we employ the vision transformer (ViT) [29] as our backbone to extract multi-layer features. We first crop the input RGB image into non-overlapping patches, followed by a linear projection to embed the patches into tokens. Then the tokens are added by position embeddings and processed by multiple transformer blocks with multiheaded self-attention. The encoder consists of 12 transformer blocks, from which we select 4 layers of tokens, evenly distributed from shallow to deep, for the following module.

(a) Pipeline

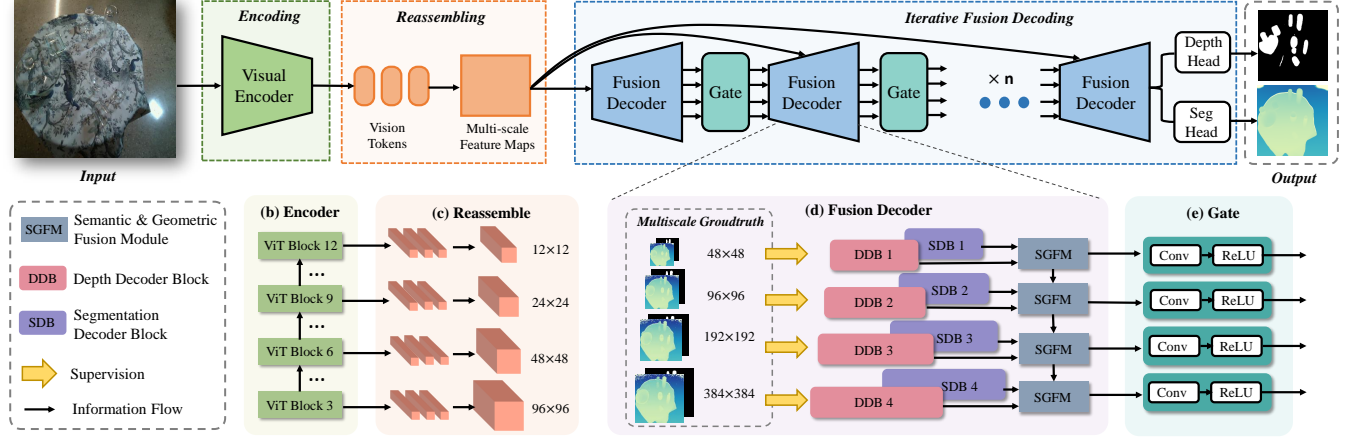


Fig. 2. **Overview of our proposed end-to-end framework.** (a) Given an RGB input, our model jointly predicts depth and segmentation mask through encoding, reassembling, and iterative fusion decoding. (b) The encoder uses ViT [29] to extract vision tokens of four layers. (c) Then in the reassemble module, the tokens are transformed into multi-scale feature maps, forming two pyramids for depth and segmentation, respectively. (d) A novel semantic and geometric fusion module is designed in the decoder for better leveraging the complementary information of both tasks. (e) The shared-weight decoder is updated iteratively by lightweight gates to gradually refine the initial results. Final predictions are obtained by two heads after the last iteration.

C. Reassemble Module

Since ViT encodes image features as tokens with the same spatial resolution, we need to convert them back to feature maps for subsequent fusion and prediction. Following DPT [30], the vision tokens are reshaped into corresponding feature maps by concatenation and projection. To fully exploit features of different levels, we represent them in a multi-scale fashion, where deeper features correspond to smaller resolutions. The results of the reassemble module are two four-layer pyramids for depth and segmentation, respectively.

D. Iterative Fusion Decoder

In the decoder, the geometric features and semantic features from the two pyramids are integrated together with our proposed fusion module. Then we iteratively refine the features from the same shared-weight decoder through gated units to obtain more fine-grained predictions.

Fusion Decoder. Due to the optical properties of transparent objects, it is particularly difficult to predict depth and segmentation independently with a single RGB image [3]. To improve the performance of both tasks, inspired by [14], we design a novel attention-based fusion module to fully exploit the complementary information of the two branches. With the two feature pyramids of depth and segmentation from the previous module, we apply semantic and geometric fusion at every layer to integrate multi-scale features. Without loss of generality, in Fig. 3 we take one layer of the features as an example. Given a depth feature $F_d \in R^{C \times H_f \times W_f}$ and a segmentation feature $F_s \in R^{C \times H_f \times W_f}$ of a certain scale, we first apply a channel attention module and a spatial attention module to successively extract meaningful cues. Attention along channel and spatial axes has been proven to be effective in learning what and where to focus [31]. Leveraging this powerful representation ability, the module can automatically learn significant semantic and geometric information implied in depth and segmentation features,

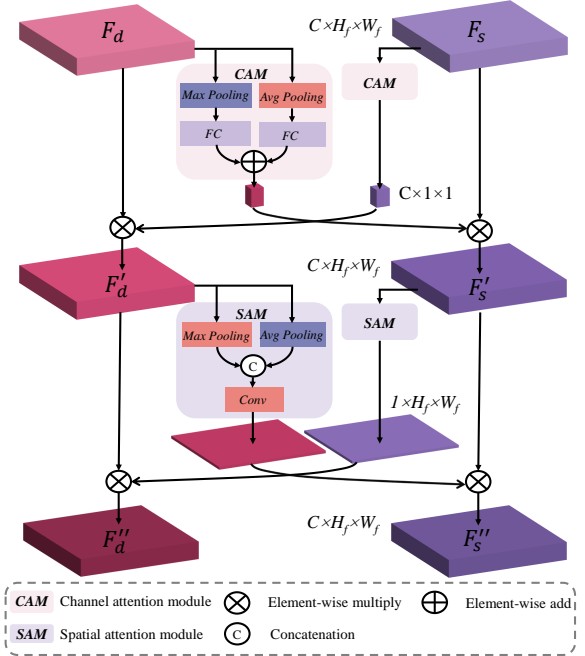


Fig. 3. **Illustration of the semantic and geometric fusion module (SGFM).** F_d and F_s represent features of a certain layer of the depth and segmentation pyramid, respectively. The two feature maps are processed along both channel and spatial dimensions to adaptively emphasize semantic and geometric information. They are then cross-multiplied to achieve the fusion.

respectively. The informative extractions then interact with each other through symmetric multiplication. Concretely, the channel attention and spatial attention are computed as:

$$CAM(F) = \sigma(FC(AP(F)) + FC(MP(F))) \quad (1)$$

$$SAM(F) = \sigma(c^{7 \times 7}[AP(F), MP(F)]) \quad (2)$$

where F represents either F_d or F_s . AP and MP denote average pooling and max pooling, respectively. FC denotes a fully connected layer. $c^{7 \times 7}$ represents a 7×7 convolution

operation. $[\cdot, \cdot]$ denotes the concatenation operation and σ is the sigmoid function.

Taking the depth feature as an example, the overall semantic and geometric fusion module is defined as:

$$F'_d = F_d \otimes CAM(F_s) \quad (3)$$

$$F''_d = F'_d \otimes SAM(F'_s) \quad (4)$$

where \otimes denotes the element-wise multiplication. F'_d and F'_s are the intermediate representation of depth and segmentation features and F''_d represents the depth feature after fusion. Segmentation features are processed in the same way.

Iterative Refinement. When faced with transparent objects, previous works with only one iteration of prediction tend to produce unclear results [12], [13]. To solve the problem, we instead propose an iterative refinement strategy to optimize depth and segmentation features in a coarse-to-fine manner. Taking the first multi-scale fusion results as the initial features, we update them repeatedly with a shared-weight decoder. The results from the previous iteration are passed to the next via lightweight gated units, which contain convolution operations and ReLU functions. The overall iterative process can be expressed as:

$$F_n = f_d(F_e, Gate(F_{n-1})) \quad (5)$$

where F_{n-1} and F_n are the set of all the multi-scale depth and segmentation features of iteration $n - 1$ and n . F_e denotes the features from the reassemble module and f_d is the function represented by the shared decoder. Based on the features from the last layer and after the last iteration, two prediction heads consisting of convolutions and interpolations are adopted to obtain the final depth map and segmentation mask. To enforce the model to learn more details about transparent objects gradually, we apply multi-scale supervision from weak to strong to each iteration. The strength of each supervision is controlled by n/N , where N is the total number of iterations and is set as 3 according to the ablation experiment.

E. Hybrid Loss Function

Our proposed model is trained end-to-end using two loss functions for depth and segmentation.

Geometric Loss. Following [9], the depth estimation loss is formulated as:

$$L_{geo} = w_d \|D - D^*\|_2 + w_g \|\nabla D - \nabla D^*\|_1 + w_n \|N_D - N_{D^*}\|_1 \quad (6)$$

where the three terms represent the L2 loss between the predicted depth D and the ground-truth depth D^* , and the L1 losses between the gradient and surface normal of D and D^* , respectively. w_d , w_g and w_n are weights and are all set as 1 in practice.

Semantic Loss. For semantic segmentation, the standard cross-entropy loss is used:

$$L_{sem} = l_{ce}(S, S^*) \quad (7)$$

where S and S^* denote the predicted and ground-truth segmentation masks, respectively.

Overall, the total loss function is formulated as:

$$L = \alpha L_{geo} + \beta L_{sem} \quad (8)$$

where α and β are two hyper-parameters empirically set to 1 and 0.1 based on their relative magnitudes. The hybrid loss function is applied to multi-scale layers of the dual-branch decoder in each iteration.

IV. EXPERIMENTS

A. Experiment Setup

Implementation Details. Our model is implemented in PyTorch and trained on an RTX 4090 GPU with a batch size of 4 for 20 epochs. For all training, we use the Adam optimizer with a learning rate of $1e-5$. The resolution of the input image is resized to 384×384 , without using any image augmentation strategies such as random flipping or rotating.

Datasets. To evaluate the effectiveness and robustness of our model, we conduct experiments on both synthetic dataset Syn-TODD [13] and real-world dataset ClearPose [32]. Syn-TODD is a photo-realistic dataset containing more than 113k image pairs with multi-task annotations, which is compatible with monocular, stereo, and multi-view methods. We follow the original paper [13] to prepare the dataset. ClearPose is a real-world dataset consisting of over 350k RGB-Depth frames. The dataset includes extreme scenarios such as heavy occlusions and non-planar configurations. We define two semantic categories, namely background and object, but note that it can be easily extended to other various classes. We follow the official setup to split the dataset into training and testing sets and the input depth map is not utilized.

Baselines. Since our model is the first monocular multi-task framework for transparent objects, we compare it with two state-of-the-art stereo and multi-view methods elaborate for transparent objects, namely SimNet [12] and MVTrans [13]. The other two baselines are InvPT [26] and TaskPrompter [28], which are designed for general multi-task dense predictions. SimNet takes stereo images as input, while MVTrans can be extended to 3 or 5 views. Both methods first construct a matching volume on the reference image through homography transformation, then perform multi-task predictions. InvPT and TaskPrompter are two state-of-the-art multi-task frameworks leveraging the interactions between different tasks with monocular images as input.

Evaluation Metrics. For depth estimation, following [3], root mean squared error (RMSE), absolute relative difference (REL), and mean absolute error (MAE) are used as standard metrics. For semantic segmentation, we use intersection over union (IoU) and mean average precision (mAP) as [13] for fair comparison. $\text{IoU} > 0.5$ is used as the threshold in computing mAP.

B. Comparison on Synthetic Dataset

As shown in Table I, we conduct experiments against other baselines on the Syn-TODD dataset. The two monocular methods are reproduced using their default settings for fair

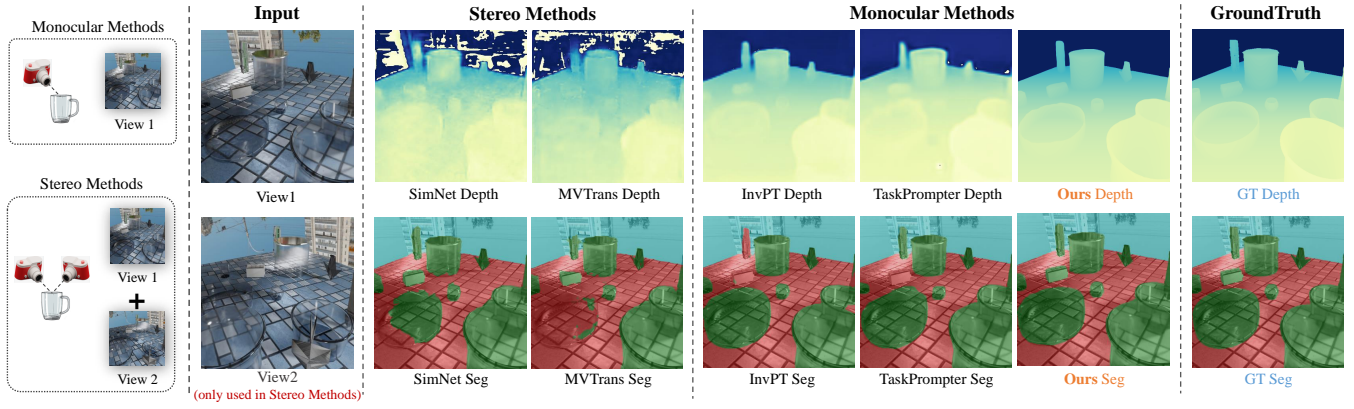


Fig. 4. **Qualitative comparison on Syn-TODD dataset** of depth and segmentation, where Seg and GT stand for segmentation and ground truth, respectively. SimNet and MVTrans take both RGB images as input, while the other methods only take the first one as input. Obviously, our predictions are far better than all other methods with only single RGB as input.

TABLE I

QUANTITATIVE COMPARISON OF SOTA MONOCULAR, STEREO, AND MULTI-VIEW METHODS ON SYN-TODD DATASET, WHERE \uparrow INDICATES THAT HIGHER VALUES ARE BETTER AND \downarrow MEANS THAT LOWER VALUES ARE BETTER. THE LAST LINE SHOWS THE PERCENTAGES BY WHICH OUR METHOD EXCEEDS THE SECOND-BEST RESULTS.

	Task	Modality	Depth			Segmentation	
			RMSE (\downarrow)	MAE (\downarrow)	REL (\downarrow)	mAP (\uparrow)	IoU (\uparrow)
InvPT [26]	general	monocular RGB	0.166	0.145	0.159	95.62	89.74
TaskPrompter [28]	general	monocular RGB	0.247	0.233	0.247	<u>96.90</u>	<u>90.50</u>
SimNet [12]	transparent	stereo RGB	1.229	1.020	0.975	48.21	50.52
MVTrans [13]	transparent	2-view RGB	0.134	0.089	0.135	84.94	79.52
MVTrans	transparent	3-view RGB	0.125	0.083	0.125	87.75	81.89
MVTrans	transparent	5-view RGB	0.124	0.080	0.117	87.24	81.30
Ours	transparent	monocular RGB	0.070	0.052	0.068	97.83	92.84
			+45.2%	+38.8%	+46.2%	+0.9%	+2.1%

TABLE II

COMPARISON OF MULTI-TASK METHODS ON THE CLEARPOSE DATASET.

	Depth			Segmentation	
	RMSE(\downarrow)	MAE(\downarrow)	REL(\downarrow)	mAP(\uparrow)	IoU(\uparrow)
InvPT	0.185	0.163	0.212	98.09	85.91
TaskPrompter	0.172	0.146	0.190	97.78	85.00
Ours	0.123	0.081	0.087	98.21	86.27

TABLE III

ABLATION STUDY ON THE SEMANTIC AND GEOMETRIC FUSION MODULE. DEPTH ONLY AND SEG ONLY INDICATE SINGLE-TASK PREDICTIONS AND W/O MEANS WITHOUT.

	Depth			Segmentation	
	RMSE(\downarrow)	MAE(\downarrow)	REL(\downarrow)	mAP(\uparrow)	IoU(\uparrow)
Ours Depth Only	0.080	0.061	0.073	-	-
Ours Seg Only	-	-	-	97.12	91.57
Ours w/o SGFM	0.087	0.064	0.076	96.17	90.28
Ours Full	0.070	0.052	0.068	97.83	92.84

comparison and the results of stereo and multi-view methods come from the original papers. Obviously, despite using only a single RGB image as input, our model significantly outperforms all other monocular and even multi-task baselines both in depth estimation and semantic segmentation, exceeding over 45% on RMSE and REL compared to the second-best method. Besides, due to the lack of inter-task

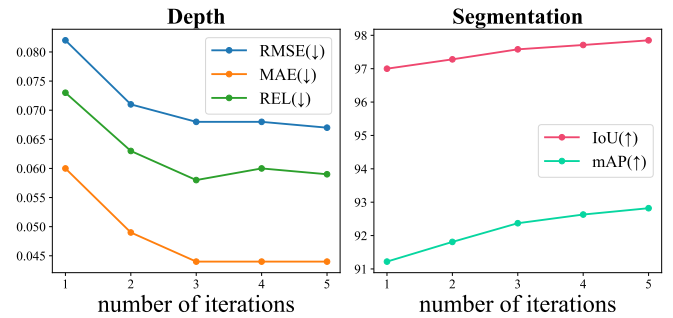


Fig. 5. Ablation studies on the iterative strategy.

communication, other methods have unbalanced performance on different tasks. For example, monocular methods perform better on segmentation than depth, while multi-view methods do the opposite. In contrast, our method performs quite well on all tasks, which can be attributed to the cross-task fusion of the complementary information.

Fig. 4 shows a qualitative test example with transparent objects. As can be seen, since transparent objects usually inconsistently refract the background and lack texture in RGB images, the baseline methods tend to make unsatisfactory predictions, resulting in large areas of missing both in depth maps and segmentation masks. However, by iteratively exchanging useful semantic and geometric information, our

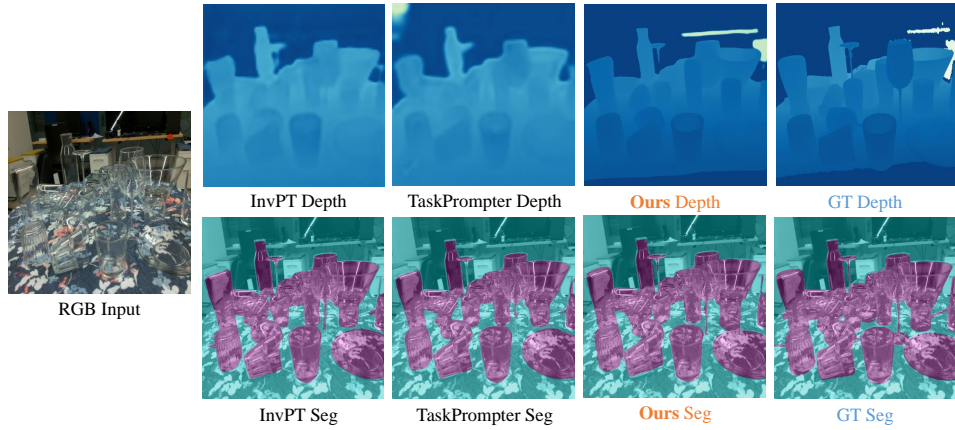


Fig. 6. **Qualitative comparison on ClearPose dataset**, where Seg and GT stand for segmentation and ground truth respectively. Although this test scene is rather challenging, our method performs quite well in both depth and segmentation compared to SOTA multi-task methods.

method produces complete and sharp-edged results.

C. Comparison on Real-world Dataset

The quantitative results on the large-scale ClearPose dataset are shown in Table II. Since this dataset does not support multi-view inputs, we only reproduce InvPT and TaskPrompter on ClearPose. Our approach consistently performs best on the real-world dataset, especially on the more challenging depth estimation task, affirming the robustness of our model.

In Fig. 6, visualizations of a complex test scene with heavy clutter and occlusion is provided to showcase the superiority of our method. Even humans find it difficult to accurately recognize and determine the geometric relations of each transparent object in this scene. In spite of achieving competitive results in 2D segmentation tasks, both InvPT and TaskPrompter lag far behind in depth estimation, producing blurry and noisy predictions. The results intuitively reveal that the multi-task baselines fail to utilize semantics to boost depth performance, and a single regression is not sufficient to obtain detailed results for transparent objects. However, our method overcomes these issues by gradually refining the multi-scale features, which allows the decoder to learn more details such as edges and surfaces.

D. Ablation Studies

Fusion Ablation. In Table III, we remove the semantic and geometric fusion module and retrain our model from scratch on Syn-TODD dataset, keeping other configurations the same. We also remove one of the depth and segmentation branches completely. The empirical results show that although predicting depth or segmentation alone is better than predicting them simultaneously, all three variants exhibit a significant drop in both depth and segmentation performance. The reason lies in that supervising the two branches independently without fusion would lead to conflict compared to single prediction, while the integration of our fusion module makes the gradient backpropagate between the two branches, which facilitates both tasks.

Iterative Strategy Ablation. To investigate the utility of our iterative strategy, in Fig. 5, we show the relationship between the number of iterations and the prediction performance of both tasks on Syn-TODD dataset. When the number of iterations is 1, the prediction process is equivalent to other models and the iterative strategy does not work. The supervision is gradually increased to ensure that the last iteration is fully supervised. It can be seen from the general trend of the curves that, as the number of iterations increases, the overall performance of both depth and segmentation improves accordingly, which demonstrates the effectiveness of our iteration strategy. The results reveal that only one regression produces suboptimal results, and by repeatedly updating the features, the model can be forced to gradually observe different details such as edges and surfaces, as humans do. Thus with an iterative strategy, our model can produce clearer dense predictions, which is especially useful for transparent objects. Considering the tradeoff between performance and memory footprint, we set the total number of iterations to 3.

V. CONCLUSIONS

In this work, we propose a monocular framework to jointly predict depth and segmentation for transparent objects. We present a semantic and geometric fusion module which can better leverage the complementary information of both tasks. An iterative strategy is also utilized to gradually refine the initial blurry results of untextured transparent objects. Experimental results demonstrate that our model outperforms state-of-the-art monocular and multi-view methods by a large margin, both on synthetic and real-world datasets.

ACKNOWLEDGMENT

This work is supported by the National Key Research and Development Program of China under Grant 2021ZD0114505, the Open Projects Program of State Key Laboratory of Multimodal Artificial Intelligence Systems under Grant No.MAIS2024112, and the Excellent Youth Program of State Key Laboratory of Multimodal Artificial Intelligence Systems.

REFERENCES

- [1] J. Jiang, G. Cao, T.-T. Do, and S. Luo, "A4t: Hierarchical affordance detection for transparent objects depth reconstruction and manipulation," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 9826–9833, 2022.
- [2] J. Jiang, G. Cao, J. Deng, T.-T. Do, and S. Luo, "Robotic perception of transparent objects: A review," *IEEE Transactions on Artificial Intelligence*, vol. 5, no. 6, pp. 2547–2567, 2023.
- [3] 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 *Proceedings of IEEE International Conference on Robotics and Automation*, 2020, pp. 3634–3642.
- [4] 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 *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 12 612–12 621.
- [5] D. Huo, J. Wang, Y. Qian, and Y.-H. Yang, "Glass segmentation with rgb-thermal image pairs," *IEEE Transactions on Image Processing*, vol. 32, pp. 1911–1926, 2023.
- [6] J. Ichnowski, Y. Avigal, J. Kerr, and K. Goldberg, "Dex-nerf: Using a neural radiance field to grasp transparent objects," in *Proceedings of Conference on Robot Learning*, 2020.
- [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 *Proceedings of Conference on Robot Learning*, 2022.
- [8] 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 *Proceedings of IEEE International Conference on Robotics and Automation*, 2023, pp. 1757–1763.
- [9] 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. 1–8, 2022.
- [10] 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 *Proceedings of European Conference on Computer Vision*, 2022, pp. 374–391.
- [11] K. Chen, S. Wang, B. Xia, D. Li, Z. Kan, and B. Li, "Tode-trans: Transparent object depth estimation with transformer," in *Proceedings of IEEE International Conference on Robotics and Automation*, 2023, pp. 4880–4886.
- [12] T. Kollar, M. Laskey, K. Stone, B. Thananjeyan, and M. Tjersland, "Simnet: Enabling robust unknown object manipulation from pure synthetic data via stereo," in *Proceedings of Conference on Robot Learning*, 2022, pp. 938–948.
- [13] 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 *Proceedings of IEEE International Conference on Robotics and Automation*, 2023, pp. 3771–3778.
- [14] J. Jiao, Y. Cao, Y. Song, and R. Lau, "Look deeper into depth: Monocular depth estimation with semantic booster and attention-driven loss," in *Proceedings of European Conference on Computer Vision*, 2018, pp. 53–69.
- [15] W. Wang and U. Neumann, "Depth-aware cnn for rgb-d segmentation," in *Proceedings of European Conference on Computer Vision*, 2018, pp. 135–150.
- [16] X. Yan, M. Sun, Y. Han, and Z. Wang, "Camouflaged object segmentation based on matching-recognition-refinement network," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–15, 2023.
- [17] E. Xie, W. Wang, W. Wang, M. Ding, C. Shen, and P. Luo, "Segmenting transparent objects in the wild," in *Proceedings of European Conference on Computer Vision*, 2020, pp. 696–711.
- [18] H. He, X. Li, G. Cheng, J. Shi, Y. Tong, G. Meng, V. Prinet, and L. Weng, "Enhanced boundary learning for glass-like object segmentation," in *Proceedings of IEEE/CVF International Conference on Computer Vision*, 2021, pp. 15 859–15 868.
- [19] H. Mei, X. Yang, Y. Wang, Y. Liu, S. He, Q. Zhang, X. Wei, and R. W. Lau, "Don't hit me! glass detection in real-world scenes," in *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3684–3693.
- [20] K. Fan, C. Wang, Y. Wang, C. Wang, R. Yi, and L. Ma, "Rfenet: Towards reciprocal feature evolution for glass segmentation," in *Proceedings of International Joint Conference on Artificial Intelligence*, 2023, pp. 717–725.
- [21] 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 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 4649–4658.
- [22] 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 *Proceedings of Conference on Robot Learning*, 2022, pp. 827–838.
- [23] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 012–10 022.
- [24] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," *Communications of the ACM*, vol. 65, no. 1, pp. 99–106, 2021.
- [25] B. P. Duisterhof, Y. Mao, S. H. Teng, and J. Ichnowski, "Residual-nerf: Learning residual nerfs for transparent object manipulation," in *Proceedings of IEEE International Conference on Robotics and Automation*, 2024, pp. 13 918–13 924.
- [26] H. Ye and D. Xu, "Inverted pyramid multi-task transformer for dense scene understanding," in *Proceedings of European Conference on Computer Vision*, 2022.
- [27] Z. Zhang, Z. Cui, C. Xu, Z. Jie, X. Li, and J. Yang, "Joint task-recursive learning for semantic segmentation and depth estimation," in *Proceedings of European conference on computer vision*, 2018, pp. 235–251.
- [28] H. Ye and D. Xu, "Taskprompter: Spatial-channel multi-task prompting for dense scene understanding," in *Proceedings of International Conference on Learning Representations*, 2023.
- [29] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," 2021.
- [30] R. Ranftl, A. Bochkovskiy, and V. Koltun, "Vision transformers for dense prediction," in *Proceedings of IEEE/CVF International Conference on Computer Vision*, 2021, pp. 12 179–12 188.
- [31] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of European Conference on Computer Vision*, 2018, pp. 3–19.
- [32] X. Chen, H. Zhang, Z. Yu, A. Opipari, and O. Chadwicke Jenkins, "Clearpose: Large-scale transparent object dataset and benchmark," in *Proceedings of European Conference on Computer Vision*, 2022, pp. 381–396.