SRNet-Trans：A Signal-Image Guided Depth Completion Regression Network for Transparent Object

Tao Tao , Hong Zheng , Jinsheng Xiao , Wenfei Wu , Jianfeng Yang *

Posted Date: 15 September 2025

doi: 10.20944/preprints202509.1153.v1

Keywords: transparent object; depth completion; signal image; self-attention; intelligent perception

Article

SRNet-Trans: A Signal-Image Guided Depth Completion Regression Network for Transparent Object

Tao Tao, Zheng Hong, Xiao Jinsheng, Wu Wenfei and Yang Jianfeng *

School of Electronic Information, Wuhan University, Wuhan, China

* Correspondence: yjf@whu.edu.cn

Abstract

Transparent objects are prevalent in various everyday scenarios. However, their reflective and refractive optical properties present significant challenges for conventional optical sensors. This difficulty makes the task of generating dense depth maps from sparse depth maps and high-resolution RGB images a critical area of research. In this paper, we introduce SRNet-Trans, a novel two-stage depth estimation framework specifically designed for transparent objects. The approach is structured into two stages, each primarily focused on leveraging semantic and depth information, respectively. In the first stage, RGB images and sparse depth maps are used to predict a relatively dense depth map. The second stage then takes the predicted depth from the first stage, along with the sparse depth map, to generate a final dense depth map. The depth information produced by the two stages is complementary, allowing for effective fusion of both outputs. To enhance the depth estimation process, we integrate a self-attention mechanism in the first stage to better capture semantic features and introduce geometric convolutional layers in the second stage to improve depth encoding accuracy. Additionally, we incorporate a global consistency-based fine depth recovery technique to further refine the final depth map. Extensive experiments on the large-scale real-world TransCG dataset demonstrate that SRNet-Trans outperforms current state-of-the-art methods in terms of depth estimation accuracy.

Keywords: transparent object; depth completion; signal image; self-attention; intelligent perception

1. Introduction

With the rapid advancement of unmanned systems, including aerial drones, ground robots, and service platforms, intelligent perception has become a cornerstone for enabling autonomous navigation, accurate positioning, and reliable task execution. In such contexts, robots must be capable of perceiving and interpreting diverse objects in their environment to ensure safety and efficiency. However, transparent objects—such as glass, plastics, and laboratory containers—pose unique challenges due to their refractive and reflective optical properties, which often disrupt conventional vision and depth-sensing mechanisms. The inability to perceive transparent objects may result in navigation errors, manipulation failures, or even safety risks. Therefore, enhancing the perception capabilities of unmanned systems toward transparent objects is of critical importance, as it directly supports applications such as collision-free navigation in cluttered scenes, robotic grasping and manipulation of fragile transparent items, and high-precision localization in safety-critical domains.

Transparent objects are ubiquitous in everyday life, with materials such as glass, plastic, and glass lids being commonly encountered. Similarly, industrial glassware—such as beakers, test tubes, and petri dishes—are integral to laboratory settings. With the increasing integration of robotics into daily activities, it is essential for robots to be capable of obtaining accurate pose information for transparent objects in their environment [1]. Depth sensing technologies, such as RGB-D cameras, play a pivotal role in achieving this. Depth maps produced by these cameras have found extensive

 doi:10.20944/preprints202509.1153.v1

applications in fields like 3D reconstruction and robotics, offering improved insight into the complex geometric details of dense scenes and fine geometric features of targets when compared to RGB images. However, transparent objects, due to their refractive and reflective properties, present a significant challenge for conventional depth sensors. These optical characteristics disrupt the geometric light path assumptions that depth sensing relies on, complicating the task of acquiring reliable depth data for such objects [2]. As a result, a hybrid approach that combines the scene geometry captured by RGB images with the sparse depth information provided by depth sensors is necessary to reconstruct a more accurate, higher-density depth map.

Depth completion for transparent objects has emerged as a challenging problem in computer vision in recent years. Given the inherent material properties of transparent objects, hardware-based solutions often struggle to address the complexities of depth recovery in general scenarios [3]. However, with the rapid advancements in neural networks and large language models, new methodologies for transparent object depth completion have emerged. Currently, approaches in this domain can be broadly classified into two categories: multi-view and single-view methods [4]. While multi-view approaches offer more comprehensive reconstruction and enhanced perception of transparent objects, they introduce additional challenges in practical scenarios. Specifically, the instability of multi-camera setups during deployment can lead to uncertainties in algorithmic results. Moreover, multi-view methods fail to leverage valuable information present in the original depth map, which may limit their adaptability, particularly in dynamic environments.

Single-view depth completion, in contrast, faces three primary challenges [5]. First, many current methods rely on the encoder-decoder structure, which is common in visual tasks, to restore depth information. However, this approach often overlooks the difficulties associated with the lack of texture features in transparent objects. Second, there is a tendency to neglect the cross-modal interaction between the shallow feature details of transparent objects and RGB-D data, leading to a loss of local details and unclear object contours in the predicted depth map. Finally, not all areas of a depth image require completion. Depth completion should be applied selectively, focusing only on regions where depth information is missing or erroneous. However, many existing methods employ full convolutional networks, treating all regions equally, which is inefficient and can result in suboptimal performance.

To address the limitations of existing approaches, this paper proposes an end-to-end deep regression network designed to achieve efficient and high-precision depth completion for transparent objects. Our method introduces a two-stage network architecture, consisting of a semantic clue-dominated stage and a depth information-dominated stage. In more detail, the semantic clue-dominated stage primarily focuses on understanding semantic information that is crucial for depth prediction. This stage emphasizes semantic cues, making the predicted depth particularly reliable around the edges of transparent objects. However, it is more sensitive to variations in color and texture. The depth information-dominated stage, on the other hand, takes both the sparse depth data and the depth predictions from the first stage as input to generate a dense depth map. While this stage typically produces more reliable depth estimates, the input sparse depth data can introduce significant noise, especially along the edges of transparent objects. Since the depth maps produced by these two stages are complementary, we perform a deep fusion of their results to achieve a more accurate final output. Additionally, we refine the fused dense depth map through a fine depth recovery process that ensures global consistency across the entire map.

The main contributions of this paper are as follows:

1.Two-Stage Network Architecture for Transparent Object Depth Completion: To fully extract features from RGB-D images of transparent objects, we propose a two-stage depth completion network tailored for semantic scenes. This architecture integrates a semantic information-guided stage and a depth information-driven stage to perform dense depth prediction, effectively leveraging and combining the cross-modal features of RGB-D data.

2.Introduction of Self-Attention for Transparent Object Depth Completion: This paper is the first to incorporate a self-attention mechanism into transparent object depth completion. The self-

 doi:10.20944/preprints202509.1153.v1

attention mechanism enables comprehensive encoding of surface normal information and edge details, significantly enhancing the performance of depth completion tasks for transparent objects.

3.Global Consistency via Scale Factor-Based Refinement: We introduce a scale factor approach to refine depth completion in the scale space, improving the global consistency of the depth map. This refinement process greatly enhances the accuracy and quality of the predicted depth map.

2. Related work

Deep image completion technology aims to utilize deep learning techniques to repair and fill in missing regions of images. It has found widespread applications in tasks such as image restoration, editing, and generation. The advent of deep learning models, including convolutional neural networks (CNNs) and generative adversarial networks (GANs), has significantly enhanced both the effectiveness and efficiency of image completion. Initially, image completion techniques were primarily based on traditional image processing methods. However, with the introduction of CNNs, researchers began leveraging the powerful feature extraction capabilities of deep learning to improve the quality of image completion.

2.1. Based on Deep Learning

Liu et al. [6] proposed the DualTransNet network, which can recover depth maps from segmentation features alone, without requiring RGB or depth map inputs. This demonstrates the effectiveness of segmentation features for depth estimation of transparent objects. Zhai et al. [7] introduced TCRNet, a transparent object depth completion network based on a cascaded refinement structure that effectively balances accuracy and real-time performance. The network utilizes a cascaded refinement mechanism during the decoding stage to iteratively refine features, thereby enhancing the accuracy of the depth information. Additionally, an attention module is incorporated to focus on the depth-related features of the transparent object regions, further improving performance. Gao et al. [8] presented a method for transparent object depth estimation based on a single RGB-D input, using a U-Net architecture with an efficient channel attention module. Despite employing a minimal number of parameters, the network significantly boosts performance. Li et al. [9] proposed a voxel-based deep learning approach for transparent object depth completion. This method leverages image features from the RGB input and valid points in the intersecting voxels derived from the point cloud. A multi-layer perceptron is used to predict the missing depth values, optimizing them under the constraint of surface normal consistency.

2.2. Based on Generative Network

Jing et al. [10] proposed a novel simulation-to-real transferable model, CAGT, which incorporates interactive embedding aggregation and geometric perception capabilities for reconstructing severely sparse depth maps of transparent objects. Pathak et al. [11] introduced the Context Encoders model, utilizing a conditional GAN architecture to enhance the visual realism of generated completion images through adversarial training. This approach effectively improves completion quality by maximizing the similarity between the generated image and the real image. The “Generative Inpainting” model, proposed by Yu et al. [12], further advances the realism of completion images by combining GANs with local context information. The adversarial training strategy within this model leads to more natural and visually coherent restoration results. Li et al. [13] investigated the effect of transparency variations on detection accuracy and proposed a detection method based on visual-tactile fusion. Their research highlighted the influence of lighting changes and the diversity of transparent object shapes on the accuracy of detection outcomes.

 doi:10.20944/preprints202509.1153.v1

3. Method

3.1. Problem Formulation

The depth completion task aims to fill in missing regions of depth measurements using the scene geometry cues provided by the corresponding RGB image. The core challenge of this task lies in efficiently fusing the geometric relationships of the monocular scene with the sparse depth data from the depth sensor—two distinct modalities—to achieve accurate depth reconstruction.

Mathematically, given a set of matched data samples, the goal is to learn a mapping function F such that $Y = F(X_{RGB}, X_{Depth})$, where $X_{RGB} \in R^{H*W}$ represents the sparse depth data, $X_{Depth} \in R^{3*H*W}$ represents the RGB image, and $Y \in R^{H*W}$ denotes the ground truth depth map.

To address this, we propose a high-performance depth completion network with a novel design that enables effective depth completion from a single RGB-D image of a transparent object. Specifically, this paper introduces a two-stage semantic scene-based depth completion algorithm tailored for transparent objects.

3.2. Network Architecture

We propose an end-to-end depth completion learning framework tailored for semantic scenes. As illustrated in the figure, the framework consists of two distinct stages: the scene geometry understanding stage, guided by semantic segmentation features, and the depth completion stage, which is primarily driven by depth information. In the backbone network, the first stage focuses on semantic information and predominantly utilizes color cues to predict relatively dense depth maps. The second stage, on the other hand, is driven by depth information, leveraging depth cues to produce even denser depth maps. The depth maps generated by these two stages are highly complementary, and we further enhance their accuracy by fusing them using confidence-based weighting. Finally, the fused depth map undergoes refinement through depth enhancement based on global consistency. The architecture of this network is designed to fully exploit and integrate the cross-modal features of RGB-D images, ensuring improved depth completion performance.

3.2.1. Semantic Information Guidance Phase Based on Self-Attention Mechanism

The first stage is the semantic scene branch, which highlights the transparent object regions based on the semantic segmentation results of the transparent object RGB-D image. This stage extracts boundary occlusion and surface normal information for depth prediction, ultimately generating a relatively dense depth map. To enhance effectiveness, the aligned sparse depth map is also incorporated for depth calibration, improving the overall depth estimation.

In this stage, the network follows an encoder-decoder architecture with a symmetrical structure: the encoder consists of one convolutional layer followed by ten residual blocks, while the decoder includes five deconvolutional layers and one convolutional layer. Depth completion involves filling in the missing gaps in a relatively sparse depth map, which can be framed as a regression problem. However, depth regression typically learns to simply copy or interpolate depth values as output. This tendency may cause the network to fall into a local minimum, where it merely copies or interpolates rather than predicting accurate depth values. To address this, we introduce a self-attention mechanism to each convolutional layer, allowing the network to focus on precise feature values at each convolution stage and output more relevant information.

 doi:10.20944/preprints202509.1153.v1

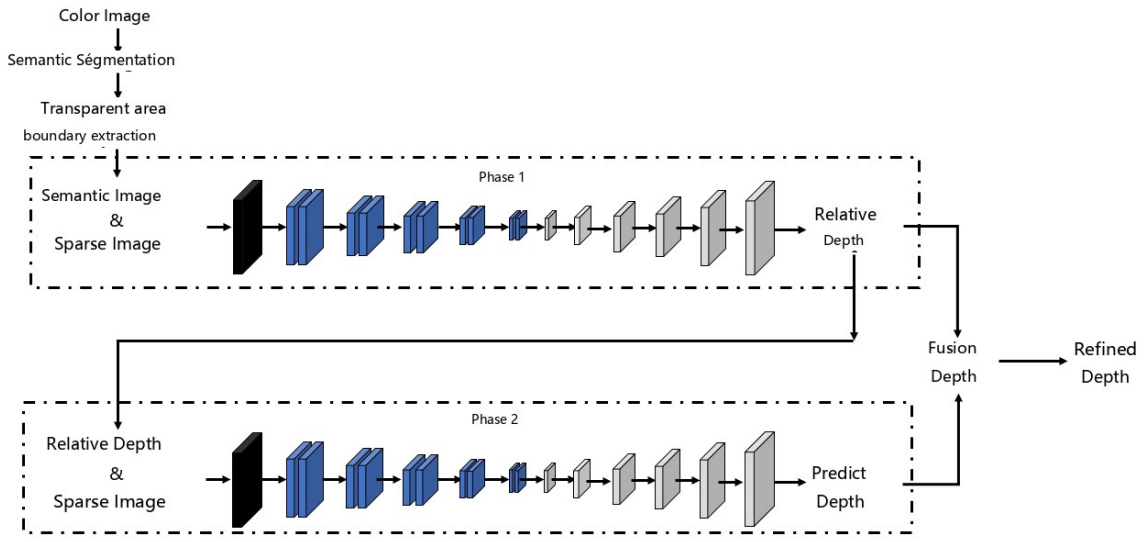

Figure 1. Overall architecture of the single-view transparent object depth completion network.

To implement this, gated convolution is employed in our network. Specifically, we define the input of a convolution block as X, the feature extraction convolution block as Conv$f$, and the gating convolution block as Conv$g$. The self-attention model can then be defined as follows:

$$Gating = sigmoid(normalization(Conv_g(X)))$$

$$Feature = sigmoid(normalization(Conv_f(X)))$$

$$Output = Gating \odot Feature$$

The normalization function Normalization (*) is used for spectral normalization, and $\odot$ represents element-wise pixel multiplication. The gating operation unique to the self-attention mechanism enables the network to dynamically select the most effective features, highlighting the semantic information within the image. As a result, the model can retain useful feature regions in the output. This convolutional network, aided by self-attention, focuses on finer image details and generates more accurate depth values.

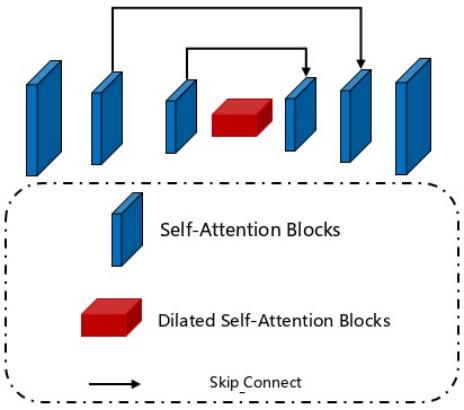

Figure 2. Self-Attention mechanismFor the self-attention network, surface normals and occlusion boundaries provide essential surface properties and texture features for transparent objects. We combine these two representations with the original sparse depth map to generate the first-stage predicted depth map, which then serves as part of the input for the second-stage network.

 doi:10.20944/preprints202509.1153.v1

3.2.2. Depth Guidance Phase Based on Geometric Convolution

The primary goal of the second stage is to predict a dense depth map by upsampling the sparse depth map. This branch also follows a similar encoder-decoder architecture. Additionally, we employ a decoder-encoder fusion strategy to integrate the semantic information-dominated features into this branch. Specifically, the decoder features from the semantic information-dominated stage are concatenated with the corresponding encoder features in the depth information-dominated branch. Furthermore, the depth prediction results from the first stage are also fed into this branch. This approach enables the fusion of color and depth modalities across multiple stages.

From a network implementation perspective, the second stage emphasizes 3D geometric cues. Building on the concept of Learning Joint 2D-3D Representations for Depth Completion, we introduce geometric convolutional layers into the encoder of this stage, replacing the conventional convolutional layers in each ResBlock to encode 3D geometric information. To enhance the convolutional layers, we incorporate the 3D position map (X, Y, Z) as additional input. The 3D position map is derived using the following formulas: $X = (u - u_0)Z/f_x,\ Y = (v - v_0)Z/f_y,\ Z = D$ where (u, v) are the pixel coordinates and (u0, v0, fx,fy) are the camera intrinsic parameters.

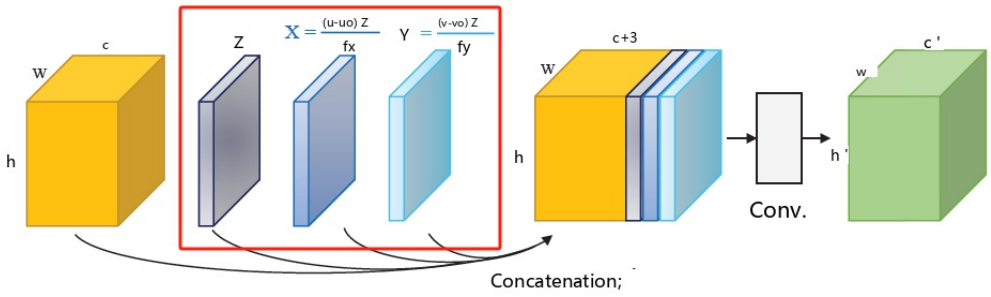

Figure 3. Geometric convolution layer.

Additionally, to better encode 3D geometric information into the depth information-dominated branch, the sparse depth map undergoes a minimum pooling operation to reduce the value of Z sufficiently.

When predicting two dense depth maps, we perform depth fusion using the following strategy:

$$\hat{Y}_f(u,v) = \frac{e^{C_{RGB}(u,v)} \bullet \hat{Y}_{RGB}(u,v) + e^{C_{Depth}(u,v)} \bullet \hat{Y}_{Depth}(u,v)}{e^{C_{RGB}(u,v)} + e^{C_{Depth}(u,v)}}$$

Here, $D1$ and $D2$ represent the depth completion results from the first and second stages, respectively, while $C1$ and $C2$ are the confidence maps corresponding to each stage.

3.2.3. Fine-Grained Depth Recovery Based on Global Consistency

Leveraging a multi-scale network based on a logarithmic space scale-independent loss function, initially proposed by Eigen, the network employs a coarse-to-fine approach for depth estimation. We make a simple assumption: adjacent pixels with similar intensities in the semantic scene segmentation image should also exhibit similar depths. This process is achieved by optimizing a weighted quadratic cost function, as described in Section 3.2.1:

$$Cost(U) = \sum_r (U(r) - \sum_{s \in N(r)} w_{rs} U(s))^2$$

Here, U represents the sparse depth to be completed, r and s refer to spatially adjacent pixels, wrs is the weight, and $N(r)$ is defined as follows:

 doi:10.20944/preprints202509.1153.v1

$$w_{rs} \propto 1+\frac{1}{\sigma_r^2}(X_{RGB}(r)-\mu_r)(X_{RGB}(s)-\mu_r)$$

where σ_r and μ_r are the mean and variance of the depth values within the r-domain window. The algorithm in this paper uses a 3×3 domain window. Additionally, the corresponding RGB image is denoted as X_{RGB}.

To enhance structural information for the depth completion task, we introduce a structure-related loss term *Ls*, which is based on the Structural Similarity Index (SSIM). SSIM evaluates the degradation of structural information, and in our task, a higher SSIM index indicates a stronger structural consistency in the completed depth map. By incorporating SSIM, we aim to guide the network to generate depth maps with better structural integrity, resulting in more refined depth completion across different scales while preserving the underlying spatial geometric structure.

3.3. Loss Function

The previous discussion demonstrated that effective depth cues can be inferred from a single transparent object image. Now, we will focus on two key aspects: the global scale of the unknown scene and the multi-scale challenges between different pixels.

To address these issues, this paper proposes an error function based on scale invariance to evaluate the accuracy of predicted depth:

$$E(Y,\hat{Y})=\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\sum_i Y_i\hat{Y}_i}{\sum_i Y_i^2}\bullet Y_i-\hat{Y}\right)^2$$

Where Y is the predicted depth, $\hat{Y}$is the true depth. Based on this formulation, we observe that multiplying Y by any non-zero scalar α results in the same error:

$$E(\alpha\bullet Y,\hat{Y})=\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\sum_i(\alpha\bullet Y_i)\hat{Y}_i}{\sum_i(\alpha\bullet Y_i)^2}\bullet(\alpha\bullet Y_i)-\hat{Y}\right)^2=E(Y,\hat{Y})$$

Thus, the error function proposed in this paper is inherently based on global scale invariance.

To simplify the calculation, we discard the terms that are independent of the predicted depth *Y* from the above formula $\frac{1}{n}\sum_{i=1}^{n}\left(\hat{Y}\right)^2$. The loss function can therefore be optimized as:

$$L=1-\frac{\left(\sum_{i=1}^{n}Y_i\hat{Y}_i\right)^2}{\left(\sum_{i=1}^{n}Y_i^2\right)\left(\sum_{i=1}^{n}\hat{Y}_i^2\right)}$$

4. Experiment

4.1. Dataset

The TransCG dataset [14] consists of 57,715 RGB images and their corresponding depth maps. It includes 51 transparent objects and approximately 200 opaque objects. All images in the dataset are captured from various real-world scenes, encompassing a total of 130 unique scenes. The objects in the dataset are randomly placed in both simple and cluttered environments, simulating real-world

 doi:10.20944/preprints202509.1153.v1

robot grasping scenarios. To maintain consistency with the original dataset's division, we use the same data split, which includes 34,191 images.

4.2. Evaluation Metrics

This paper continues to utilize the evaluation metrics from previous works [15,16], employing them to compare the performance of the networks.

(1) Root Mean Squared Error (RMSE): We calculate RMSE to evaluate the error between the predicted depth and the ground truth, which can be calculated by:

$$RMSE = \sqrt{\frac{1}{Y}\sum_{y \in Y} \left\| y - \hat{y} \right\|^2}$$

(2) Absolute Relative Difference (REL): We calculate REL to indicates the mean absolute relative difference, which can be calculated by:

$$REL = \frac{1}{Y}\sum_{y \in Y} \frac{\left| y - \hat{y} \right|}{\hat{y}}$$

(3) Mean Absolute Error (MAE): We use MAE to calculate the mean absolute error between estimated depth and ground truth, which can be calculated by:

$$MAE = \frac{1}{Y}\sum_{y \in Y} \left| y - \hat{y} \right|$$

(4) Threshold: We use the threshold to calculate the percentage of pixels with predicted depths, which can be calculated by:

$$\max\left(\frac{y}{\hat{y}}, \frac{\hat{y}}{y}\right) < threshold$$

In this paper, we set the threshold with 1.05, 1.10, and 1.25.

In the above formula, y represents a pixel in the predicted depth map$Y$, and$\hat{y}$ represents the corresponding pixel in the ground-truth depth map $\hat{Y}$.

4.3. Ablation Experiment

We first conducted a series of experiments to validate the effectiveness of the specialized design components proposed in this paper, including the two-stage backbone architecture, the incorporation of a self-attention mechanism, and deep refinement based on scale factors.

Effectiveness of the Two-Stage Backbone Structure: We propose four variants of the backbone, differentiated by whether the sparse depth map is input into the semantic guidance stage and whether the first-stage depth prediction is used as input for the depth-dominant branch. The performance of these variants, labeled M1 to M4, is shown in Table 1. The results indicate a significant performance improvement when the relative depth input (SG-Input relative depth) benefits from semantic guidance assistance and depth-dominant support. Additionally, we explore another backbone variant (M5), inspired by FusionNet and DeepLiDAR, which generates an additional guidance map from the first stage to assist the second stage. The results suggest that this extra guidance map is unnecessary and even slightly detrimental to performance. Figure 4 illustrates some typical examples.

Table 1. Ablation test results of the effectiveness of the two-stage backbone structure.

Models	SG-Input Sparse depth	DD-Input relative depth	Guidance Map	Self-attention	Refined	RMSE↓	MAE↓	REL↓
M1						0.0302	0.0176	0.0731

 doi:10.20944/preprints202509.1153.v1

M2	√					0.0299	0.0172	0.0532
M3		√				0.0295	0.0157	0.0517
M4	√	√				0.0291	0.0153	0.0515
M5	√	√	√			0.0291	0.0155	0.0619
M4+SA	**√**	**√**		**√**		**0.0190**	**0.0113**	**0.0254**
M4+SA+C1	**√**	**√**		**√**	**√**	**0.0138**	**0.0107**	**0.0155**

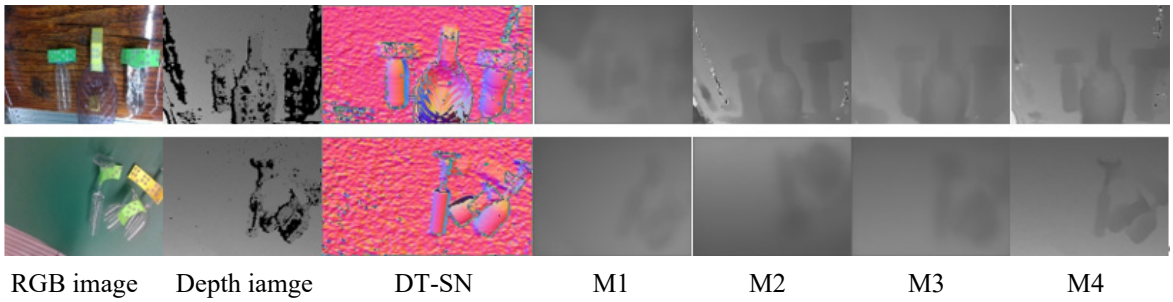

Figure 4. Results of the two-stage backbone structure validity ablation experiment.

Effectiveness of the Self-Attention Mechanism: As shown in Table 1, the inclusion of the self-attention convolutional layer significantly improves the performance of the backbone network, particularly in terms of RMSE. When the deep refinement module (Re) is added, the final model (M4+SA+Re) achieves superior detection accuracy, as indicated in the last row of the table.

Figure 5 presents a typical example to illustrate the differences between the models. The model with the self-attention convolutional layer demonstrates superior depth inference, particularly when the color features of foreground transparent objects are obscured by the background color. The first two rows of Table 2 examine the impact of the self-attention mechanism on full-depth results. Indeed, the self-attention mechanism provides a significant performance boost over the FCN model. For comparison, we use ResNet18, which has similar parameters to the classic FCN. The results indicate that this improvement stems from the network's ability to attend to convolutional features, allowing the model to focus on critical areas and key features. In this context, the self-attention mechanism enhances the model's ability to learn and retain geometric information.

SSIM Loss Function Based on Global Consistency: By introducing a smaller weight for the SSIM loss during optimization, the self-attention network learns to balance structural information without significantly affecting RMSE and delta percentage. After incorporating the SSIM loss, the SSIM score increased by 15.3%, demonstrating that the network successfully generates more accurate depth map values.

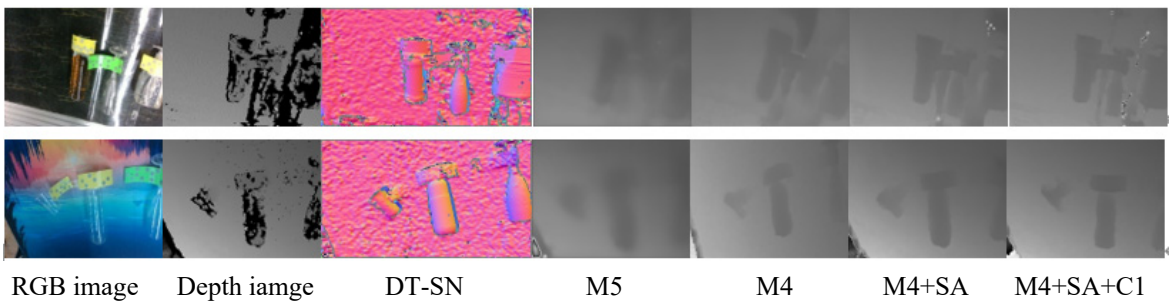

Figure 5. Self-attention mechanism effectiveness ablation experiment results.

Table 2. SSIM loss function effectiveness ablation experiment results.

Model	RMSE↓	Mean↓	SSIM↑	1.05↑	1.10↑	1.25↑
SA	0.1095	0.400	0.673	79.13	93.60	98.57

 doi:10.20944/preprints202509.1153.v1

SA+SSIM	0.1096	0.407	0.776	88.47	92.45	95.49
SA+SSIM+GC	**0.0138**	**0.392**	**0.799**	**90.41**	**97.11**	**99.72**

4.4. Comparison with SOTA

Table 3 presents the full quantitative performance of our model, along with comparisons to the top five published or archived papers. The experiments were conducted using Python 3.8, PyTorch 2.1.2, Ubuntu 20.04, and a single 4090 GPU. The results reveal significant improvements in RMSE, which is the primary evaluation metric.

Table 3. Experimental results compared with SOTA.

Model	RMSE↓	MAE↓	REL↓	1.05↑	1.10↑	1.25↑
ClearGrasp [17]	0.0540	0.0370	0.0831	50.48	68.68	95.28
LIDF-Refine [18]	0.0393	0.0150	0.0340	78.22	94.26	99.80
TranspareNet [19]	0.0361	0.0134	0.0231	88.45	96.28	99.42
TransCG [14]	0.0182	0.0123	0.0270	83.76	95.67	99.71
TODE-Trans [20]	0.0271	0.0216	0.0487	64.24	86.98	99.51
TCRNet [7]	0.0170	0.0109	0.0200	88.96	96.94	**99.87**
Ours	**0.0138**	**0.0107**	**0.0155**	**90.41**	**97.11**	99.72

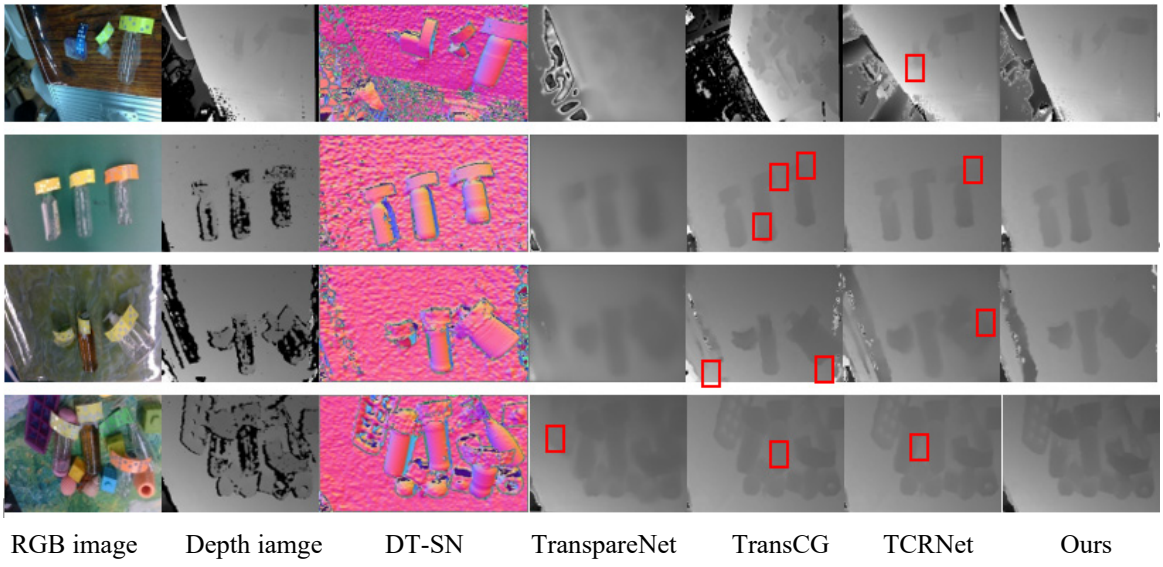

Figure 6. Experimental results compared with SOTA.

5. Conclusions

This paper introduces a novel approach to depth completion for transparent object RGB-D images. We designed a two-stage depth completion network: the semantic information-guided stage and the depth-information-dominant stage, which effectively realizes the cross-modal fusion of RGB-D images for accurate depth completion. Additionally, we are the first to apply the self-attention mechanism to depth completion of transparent objects, further enhancing the network's performance. Finally, we implemented refined depth completion processing, leading to substantial improvements in the predicted depth map. The effectiveness and superiority of our approach were validated through extensive comparison with state-of-the-art (SOTA) methods.

Funding: This work was funded by Pingyang, Zhejiang Province of China (No. 250071494).

 doi:10.20944/preprints202509.1153.v1

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data available on request from the authors.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Shi J, Yong A, Jin Y, et al. Asgrasp: Generalizable transparent object reconstruction and 6-dof grasp detection from rgb-d active stereo camera[C]//2024 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2024: 5441-5447.
2. Jing X, Qian K, Vincze M. CAGT: Sim-to-Real Depth Completion with Interactive Embedding Aggregation and Geometry Awareness for Transparent Objects[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2025.
3. Ummadisingu A, Choi J, Yamane K, et al. Said-nerf: Segmentation-aided nerf for depth completion of transparent objects[C]//2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2024: 7535-7542.
4. Jin Y, Liao L, Zhang B. Depth Completion of Transparent Objects Based on Feature Fusion[C]//2024 4th International Conference on Artificial Intelligence, Virtual Reality and Visualization. IEEE, 2024: 95-98.
5. Meng X, Wen J, Li Y, et al. DFNet-Trans: An end-to-end multibranching network for depth estimation for transparent objects[J]. Computer Vision and Image Understanding, 2024, 240: 103914.
6. Liu B, Li H, Wang Z, et al. Transparent Depth Completion Using Segmentation Features[J]. ACM Transactions on Multimedia Computing, Communications and Applications, 2024, 20(12): 1-19.
7. Zhai D H, Yu S, Wang W, et al. Tcrnet: Transparent object depth completion with cascade refinements[J]. IEEE Transactions on Automation Science and Engineering, 2024.
8. Gao J J, Zong Z, Yang Q, et al. An Enhanced UNet-based Framework for Robust Depth Completion of Transparent Objects from Single RGB-D Images[C]//2024 7th International Conference on Computer Information Science and Application Technology (CISAT). IEEE, 2024: 458-462.
9. Li J, Wen S, Lu D, et al. Voxel and deep learning-based depth complementation for transparent objects[J]. Pattern Recognition Letters, 2025.
10. Jing X, Qian K, Vincze M. CAGT: Sim-to-Real Depth Completion with Interactive Embedding Aggregation and Geometry Awareness for Transparent Objects[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2025.
11. Pathak D, Krahenbuhl P, Donahue J, et al. Context encoders: Feature learning by inpainting[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 2536-2544.
12. Yu J, Lin Z, Yang J, et al. Generative image inpainting with contextual attention[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 5505-5514.
13. Li S, Yu H, Ding W, et al. Visual–tactile fusion for transparent object grasping in complex backgrounds[J]. IEEE Transactions on Robotics, 2023, 39(5): 3838-3856.
14. Fang H, Fang H S, Xu S, et al. Transcg: A large-scale real-world dataset for transparent object depth completion and a grasping baseline[J]. IEEE Robotics and Automation Letters, 2022, 7(3): 7383-7390.
15. J. Xiao, Y. Wu, Y. Chen, et. al., "LSTFE-Net: Long Short-Term Feature Enhancement Network for Video Small Object Detection," 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 2023, pp. 14613-14622.
16. J. Xiao, H. Guo, J. Zhou, et. al., Tiny object detection with context enhancement and feature purification, Expert Systems with Applications, 2023, vol 211, 118665.
17. Sajjan S, Moore M, Pan M, et al. Clear grasp: 3d shape estimation of transparent objects for manipulation[C]//2020 IEEE international conference on robotics and automation (ICRA). IEEE, 2020: 3634-3642.

 doi:10.20944/preprints202509.1153.v1

18. Zhu L, Mousavian A, Xiang Y, et al. RGB-D local implicit function for depth completion of transparent objects[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 4649-4658.
19. 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 Proc. Conf. Robot Learn., 2022, pp. 827–838
20. Chen K, Wang S, Xia B, et al. Tode-trans: Transparent object depth estimation with transformer[C]//2023 IEEE international conference on robotics and automation (ICRA). IEEE, 2023: 4880-4886.