Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

TADNet: Transparency-Aware Feature Aggregation with Structural Enhancement for Transparent Object Depth Completion

Xinyang Cai
Fuzhou University
Yiwen Qi
Fuzhou University
Hedan Liu
liuhedan@xujc.com

Xiamen University Tan Kah Kee College

Research Article

Keywords: Depth completion, transparent objects, transparency-aware, structural enhancement

Posted Date: September 14th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9690581/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

TADNet: Transparency-Aware Feature Aggregation with Structural Enhancement for Transparent Object Depth Completion

Xinyang Cai[1], Yiwen Qi[1,2*†], Hedan Liu[3,4*†]

[1]School of Advanced Manufacturing, Fuzhou University, Quanzhou, 350116, Fujian, China.

[2]School of Electrical Engineering and Automation, Fuzhou University, Fuzhou, 350108, Fujian, China.

[3]School of Information Science and Technology, Xiamen University Tan Kah Kee College, Zhangzhou, 363123, Fujian, China.

[4]Key Laboratory of Intelligent Manufacturing Equipment and Industrial Internet Technology, Fujian Provincial Universities, Fujian, China.

*Corresponding author(s). E-mail(s): qiyiwen@fzu.edu.cn; liuhedan@xujc.com;
Contributing authors: 248527104@fzu.edu.cn;
†These authors contributed equally to this work.

Abstract

Accurate depth perception of transparent objects is crucial for robotic grasping and scene understanding, but RGB-D sensors experience serious depth noise, missing measurements, and geometric distortion when measuring transparent objects due to complex optical phenomena such as refraction and reflection. Existing depth completion models for transparent objects cannot balance global context with local structural modeling, which limits the quality of depth reconstruction in transparent regions. To address this problem, this paper proposes TADNet, a hierarchical encoder-decoder network for transparent object depth completion. First, a cascaded Swin-Transformer encoder is used to capture long-range contextual dependencies in multi-scale feature representations, providing reliable inference cues for regions with corrupted or missing depth. Second, a Transparency-Aware Attention (TAA) module is introduced to optimize global context modeling in transparent regions through region-differentiated feature aggregation guided by pixel-level transparency masks, while preserving local geometric details in non-transparent regions. At the same time, a Structural Feature Enhancement (SFE) module is designed to optimize the feature representation of structures through parallel attention, thereby promoting the reconstruction of depth discontinuities and fine-grained geometric details. Finally, a convolutional decoder with densely connected upsampling units is employed to progressively recover dense depth maps with anti-aliasing properties. Extensive experiments on the ClearGrasp and TransCG datasets demonstrate that the proposed method achieves competitive performance compared with existing approaches. The results indicate that TADNet effectively improves depth completion accuracy for transparent objects.

Keywords: Depth completion, transparent objects, transparency-aware, structural enhancement.

1 Introduction

Transparent objects are widely encountered in daily environments, such as glass containers, plastic bottles, and transparent tools. Accurate perception of transparent objects is essential for many robotic applications, including grasping [1, 2] and human–robot interaction [3, 4]. However, transparent materials pose significant challenges for RGB-D sensors, as light refraction and reflection often lead to severe depth noise, missing measurements, and unreliable geometric cues [5]. Consequently, depth maps captured in scenes containing transparent objects are often incomplete or highly distorted. As illustrated in Fig. 1, these primarily manifest as two phenomena: background penetration due to light refraction, and missing measurements caused by specular reflections.

To address this problem, recent studies [6–13] have explored depth completion techniques based on geometric priors, implicit neural representations, CNNs, and Transformer architectures. Although these learning-based methods have shown promising results, it is still challenging to extract the geometric information from transparent objects. The drawback is that many models cannot trade off global context for local structural modeling, which limits depth reconstruction quality in transparent regions.

Transformers [14] exhibit strong capabilities in capturing long-range dependencies through their self-attention mechanism, while convolutional operations remain effective for reconstructing local spatial details. Motivated by these complementary characteristics, we propose TADNet, a transparent object depth completion network designed to improve depth reconstruction in transparent scenes. The proposed framework adopts a hierarchical encoder–decoder architecture. Specifically, a Swin-Transformer encoder [15] is employed to extract rich global contextual representations from the input features, while a convolutional decoder progressively recovers dense depth maps and refined details.

To better handle the properties of transparent objects, we introduce a transparency-aware attention (TAA) module, which performs region-specific feature fusion through pixel-level transparency masks. For transparent regions, we aggregate scene information using global context blocks; while for non-transparent regions, we utilize dense convolutional blocks [16] to preserve local details, thereby enabling depth reconstruction. In addition, we design a structural feature enhancement (SFE) module, which enhances the representation of local structural transitions in the feature space, thereby assisting in the reconstruction of depth discontinuities and geometric details during decoding.

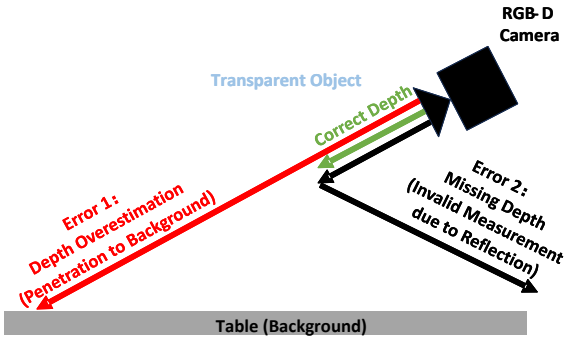


Fig. 1 Illustrations of two common depth perception errors for transparent objects

Experiments on different datasets demonstrate that our proposed method achieves competitive performance compared to state-of-the-art methods. Experimental results show that TADNet effectively improves depth completion accuracy in scenes containing transparent objects and exhibits good generalization in real-world scenarios.

In summary, the main contributions of this paper can be summarized as follows:

- We propose TADNet, a hierarchical encoder–decoder network for transparent object depth completion. This network employs a Swin-Transformer encoder to extract multi-scale contextual features from the RGB-D input and combines this with CDCU units to progressively reconstruct a dense depth map. Through dense feature reuse and anti-aliasing upsampling, this network effectively preserves structural details during depth recovery.
- A TAA module is designed to address the inherent differences in feature modeling requirements between transparent and non-transparent regions. This module performs region-differentiated feature aggregation to achieve scene-level contextual information fusion and local feature optimization for non-transparent regions.
- An SFE module is designed to optimize the feature representation of structures. This module enhances the sensitivity of the feature space to depth discontinuities and local geometric changes through a parallel attention mechanism.
- Experiments conducted on the ClearGrasp and TransCG datasets demonstrate that the proposed method is competitive with existing methods.

2 Related Work

2.1 Transparent Objects Perception

Transparent object perception has attracted widespread attention in computer vision and robotics due to its importance in robot manipulation and scene understanding. However, transparent materials introduce complex optical phenomena such as refraction and reflection, which often lead to severe distortion or measurement gaps in RGB-D sensors. This has become a core bottleneck limiting downstream tasks such as 3D reconstruction, depth completion, segmentation, pose estimation, and robot grasping.

With the development of deep learning, data-driven methods are increasingly being used for transparent object perception. ClearGrasp [6] is a representative work for depth completion of transparent objects. It estimates surface normals and object masks from RGB images and then reconstructs the depth map through a global optimization process. Although this method has shown good results, the processing speed of this optimization-based reconstruction limits its applicability to real-world scenarios.

Subsequent research has explored other learning-based strategies. For example, LIDF-Refine [7] introduces implicit neural representations to model the geometry of transparent objects, enabling the reconstruction of continuous surfaces from sparse observation data. Other methods attempt to directly improve depth estimation using deep neural networks. For instance, DFNet [8] employs the U-Net framework to enhance depth inference for transparent regions, while TODE-Trans [11] integrates a Transformer-based architecture to improve contextual reasoning during depth reconstruction. DistillGrasp [17], based on knowledge distillation, utilizes a teacher-student network to achieve excellent performance in both accuracy and speed for transparent depth completion. The aforementioned methods are all single-view approaches, while multi-view methods [18–20] have a performance advantage because they can aggregate scene features from multiple perspectives, and therefore are also widely studied. Dex-NeRF [18] uses the NeRF [21] method to recover depth information of transparent objects for grasping tasks. TranSplat [20] uses a latent diffusion model to generate surface embeddings and performs joint Gaussian optimization using both the surface embeddings and the input RGB image in the Gaussian Splatting pipeline, ultimately achieving accurate and dense depth reconstruction. Multi-view methods rely on information from different perspectives in the scene, which makes them less generalizable to new scenes than single-view-based methods. Therefore, single-view methods are more often deployed in real-world tasks.

Despite these advances, accurate perception of transparent object geometry remains difficult due to unreliable depth measurements and uncertainty of visual information. Many existing solutions either rely on optimization or struggle to capture global context and fine-grained structural information simultaneously.

2.2 Depth Completion and Estimation

Depth estimation and completion are major computer vision tasks that aim to extract dense geometric depth information of a scene from image data, providing core geometric support to robot perception, autonomous driving, and 3D reconstruction. In the existing literature, there are three types based on the input format and task goals.

The first category is monocular depth estimation [22, 23], which uses a single RGB image to predict a dense depth map. Monocular depth estimation technology has advantages such as low hardware requirements and flexible deployment, and has become the most widely adopted approach for passive depth perception in general vision scenarios. The second category is sparse depth completion [24, 25], which uses RGB images and sparse depth points as input to recover dense depth maps. It is mainly used in autonomous driving scenarios that fuse LiDAR and cameras. Sparse depth completion has become a standard technology for multi-sensor fusion depth perception in outdoor navigation and autonomous driving scenarios. The third category is dense depth completion with noisy RGB-D [6, 26], using paired RGB images and dense depth maps with noise and missing values as input, to repair and complete depth information. This is the main emphasis of this paper. Such methods are particularly well-suited for specific scenarios, such as indoor robotic applications operating in environments containing transparent or reflective objects.

3 Method

3.1 Overview

Fig. 2 illustrates the overall architecture of the proposed TADNet. Our model employs an end-to-end encoder–decoder design, taking an RGB-D image $I \in \mathbb{R}^{4\times H\times W}$, obtained by concatenating an RGB image $C \in \mathbb{R}^{3\times H\times W}$ and the original depth image $D \in \mathbb{R}^{1\times H\times W}$ along the channel dimension, as input. The output is a complete dense depth map with the same resolution as the input.

The model first feeds the input RGB-D image into the encoder to extract multi-scale global contextual features. The encoder uses a 4-level Swin-Transformer architecture, outputting feature maps at resolutions of 1/2, 1/4, 1/8, and

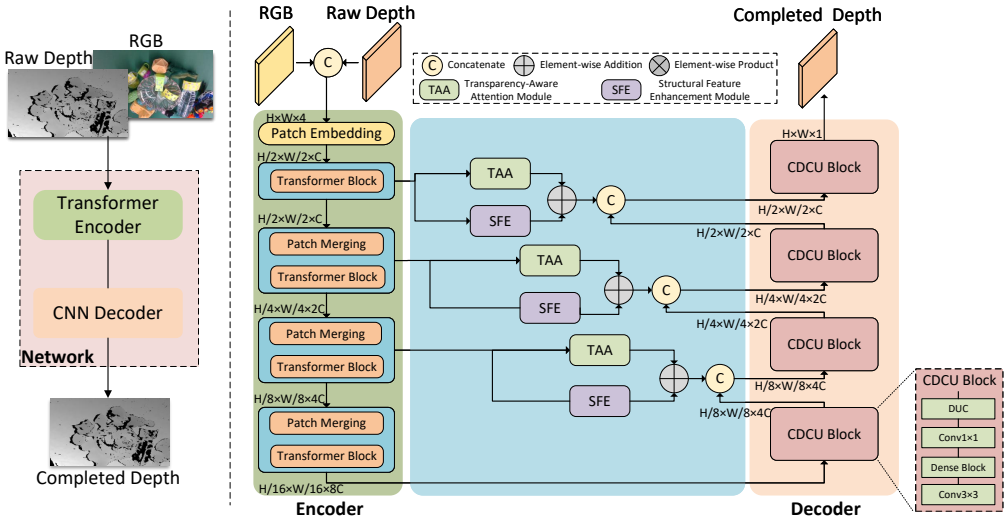


Fig. 2 Architecture of our proposed method TADNet. The network adopts a hierarchical encoder–decoder architecture, incorporating the TAA module and the SFE module for improved depth completion in transparent scenes

1/16 of the original input. For each stage of the encoder's output feature map, inputs are simultaneously processed by two modules: First, the feature map is fed into a TAA module for region-specific global-local feature fusion. Guided by pixel-level transparency masks, this module emphasizes global context modeling of transparent regions while preserving local details in non-transparent regions. Second, the same encoder feature map is fed into the SFE module to optimize feature representations related to structural changes. After processing by the TAA module and SFE module, the enhanced multi-scale features are concatenated with the corresponding layer features of the decoder via skip connections. The concatenated features are then fed into the next layer of the decoder for further feature upsampling and fusion. The decoder consists of cascaded CDCU units that progressively upsample the low-resolution fused features to the original image resolution to obtain a complete dense depth map.

In the following sections, we will detail the design of each core component and the loss function used for model training.

3.2 Swin-Transformer Encoder

Considering that depth perception of transparent objects is highly dependent on the background and lighting environment, we believe that extracting global contextual features is key to improving completion accuracy. Therefore, we introduce the Swin-Transformer in the encoder section. Specifically, the input RGB-D image is first split into several non-overlapping 2×2 image patches and projected to a C=24 dimensional space. Subsequently, these features enter a feature extraction pipeline consisting of four cascaded stages. Each stage contains an image patch merging layer and multiple Swin-Transformer modules: the former halves the resolution and doubles the dimension through downsampling, while the latter is responsible for feature modeling. The Swin-Transformer modules alternately use window self-attention mechanism (W-MSA) and shifted window self-attention mechanism (SW-MSA), achieving global context modeling with linear computational complexity relative to image resolution. The encoder finally produces hierarchical features F_1–F_4 at resolutions from $H/2$ to $H/16$ with channel dimensions from 24 to 192. These multi-scale features are passed to the decoder via skip connections and input into the TAA and SFE modules.

3.3 Decoder with CDCU Units

In the decoder design, we use a recovery network consisting of four cascaded CDCU units. The goal is to enable the model to progressively reconstruct the complete depth map from the low-resolution features output by the encoder. Each CDCU unit contains a pair of 3×3 convolutional layers, a dense block, and a dense upsampling convolutional block (DUC). This dense feature reuse is primarily to fully extract the feature values of each layer and address the vanishing gradient problem. Regarding upsampling, we abandon ordinary transposed convolutions and adopt DUC blocks because transposed convolutions are prone to checkerboard artifacts. The DUC first expands

the channel count by a factor of 2^2 using 1×1 convolutions, and then restores the high resolution of the feature map through pixel rearrangement. The advantage of this approach is that it maintains spatial continuity and eliminates aliasing. The input features are first processed in the CDCU unit, then concatenated with the features optimized by the corresponding encoder's TAA module and SFE module, and finally passed to the next stage decoder. This design ultimately ensures a richly detailed and complete depth map output.

3.4 Transparency-Aware Attention Module (TAA)

Since transparent objects and their surrounding areas in an image have different spatial requirements for feature extraction, transparent areas need to use large-scale spatial information to understand the depth variation pattern, while surrounding areas need to preserve details with precise local features. We propose the TAA module, which uses pixel-level transparency masks to guide the fusion of region-differential features. The TAA module structure is shown in Fig. 3.

This module first divides the multi-scale features output by the encoder into two dedicated processing streams using a pixel-level binary transparency mask $M \in \mathbb{R}^{H\times W}$. Here, the mask is downsampled to match the resolution of each encoder feature map. During training, we use real masks to provide noise-free supervision, ensuring that the model can capture more accurate feature maps at transparent boundaries. During inference, our framework can accept manually annotated masks or masks generated by pre-trained segmentation models. For the transparent region branch, we introduce a Global Context (GC) Block [27]. The GC Block effectively models the global context by computing query-independent attention maps, enabling the aggregation of environmental information across the entire scene, utilizing features of surrounding stable objects to repair the depth information of transparent surfaces. The global context modeling process can be described as follows:

$$F_{tr} = F_{in} + W_{v2}\,\mathrm{ReLU}\left(\mathrm{LN}\left(W_{v1}\sum_{j=1}^{N_p}\alpha_j x_j\right)\right), \tag{1}$$

where $\alpha_j = \exp(w_k x_j)/\sum_{m=1}^{N_p}\exp(w_k x_m)$ is the spatial attention weight; F_{in} and x_j represent the input feature map and the j^{th} pixel's feature vector. w_k are the convolutional parameters used to calculate the spatial attention weights, W_{v1} and W_{v2} represent the 1×1 convolutional weights used for channel compression and recovery in the bottleneck structure, respectively, and LN represents layer normalization.

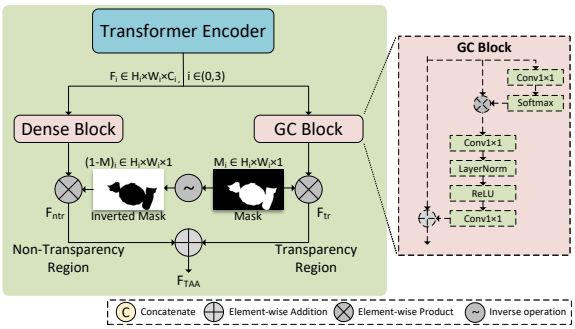


Fig. 3 Structure of the proposed TAA module, which performs region-aware feature aggregation to enhance feature representation in transparent regions

For the non-transparent region branch, we use the same dense convolutional blocks as the CDCU unit. This branch focuses on preserving local details and fine boundaries, maintaining the structural integrity of the scene. The dense convolutional blocks achieve feature reuse through channel concatenation, producing the non-transparent branch output F_{ntr}.

The outputs of the two branches are multiplied by the transparent mask M and its inverted mask $1-M$, respectively, and then summed to obtain the feature F_{TAA}:

$$F_{\mathrm{TAA}} = M \otimes F_{\mathrm{tr}} + (1-M)\otimes F_{\mathrm{ntr}}, \tag{2}$$

where $\otimes$ represents element-wise multiplication, F_{tr} is the output of the transparent region branch, and F_{ntr} is the output of the non-transparent region branch, which allows the network to trade off global context and local details at object boundaries.

3.5 Structural Feature Enhancement Module (SFE)

Since refractive blurring and sharp depth changes lead to depth estimation errors around structural transition regions, we propose the SFE module that optimizes feature representations of structural transitions.

As shown in Fig. 4, the SFE module receives the feature map $F_{in} \in \mathbb{R}^{C\times H\times W}$ from the encoder. To improve the discriminative power of local structural features while maintaining computational efficiency, the module first performs feature preprocessing using a depthwise separable convolutional block. Its calculation is defined as follows:

$$F_{mid} = \mathrm{LR}(\mathrm{BN}(\mathrm{PW}(\mathrm{LR}(\mathrm{BN}(\mathrm{DW}\,(F_{in})))))), \tag{3}$$

where DW and PW represent depthwise convolution and pointwise convolution, respectively, and LR is the LeakyReLU activation function.

Then, the feature F_{mid} simultaneously enters the parallel Channel Attention (CAM) and Spatial Attention (SAM) branches. In the CAM branch, to capture complementary channel statistics, we simultaneously employ global average pooling and global max pooling, and extract channel weights through a shared lightweight convolutional network.

$$F_{ca} = F_{mid} \otimes \sigma(\mathrm{FC}(\mathrm{Swish}(\mathrm{FC}(z_{ca})))), \quad (4)$$

where $z_{ca} = \mathrm{GAP}(F_{mid}) + \mathrm{GMP}(F_{mid})$ aggregates channel statistics via global average and max pooling. σ and $\otimes$ denote Sigmoid and channel-wise scaling. This process allows the model to adaptively emphasize feature channels that are more sensitive to structural changes.

In the SAM branch, considering the impact of illumination variations, we introduce a multi-statistical feature fusion strategy. We simultaneously extract four statistical features across the channel dimensions: mean, maximum, minimum, and sum, and concatenate them into a four-channel feature map F_{st}:

$$\begin{cases} F_{st} = [\mathrm{mean}; \max; \min; \mathrm{sum}](F_{mid}) \in \mathbb{R}^{4\times H\times W}, \\ F_{sa} = F_{mid} \odot \sigma(\mathrm{C}_{1\times 1}(\mathrm{Swish}(\mathrm{C}_{7\times 7}(F_{st})))), \end{cases} \quad (5)$$

where F_{st} concatenates four channel-wise statistics; $\mathrm{C}_{k\times k}$ is a $k \times k$ convolution; $\odot$ denotes spatial-wise weighting.

Finally, the outputs of the two branches are collaboratively combined element-wise: $F_{SFE} = F_{ca} + F_{sa}$.

The SFE and TAA module are applied in parallel between each encoder–decoder layer. The outputs of the two branches are integrated by element-wise summation to obtain the refined feature F_{fused}:

$$F_{fused} = F_{TAA} \oplus F_{SFE}, \quad (6)$$

this allows for the complementary fusion of transparency-aware semantic cues and structure feature representations.

3.6 Loss Function

To train TADNet, we employ a loss function focused on depth completion for transparent object regions, consisting of depth regression loss and surface normal smoothing loss. The loss is defined as:

$$\mathcal{L} = \mathcal{L}_d + \beta \mathcal{L}_{smooth}, \quad (7)$$

where β=0.1 is the weight parameter, L_d is the depth loss, and L_{smooth} is the cosine distance of the surface normal calculated from the predicted depth and the true depth. The L_d loss function is used to minimize the squared difference between the predicted depth and the true depth. Its calculation formula is:

$$\mathcal{L}_d = \left\| \hat{\mathcal{D}} - \mathcal{D}^* \right\|^2, \quad (8)$$

where $\hat{\mathcal{D}}$ and D^* represent the predicted depth and the true depth, respectively.

L_{smooth} is the cosine similarity loss of the surface normals. This loss measures the cosine distance between the predicted depth surface normal vector and the true depth surface normal vector. Its calculation formula is:

$$\mathcal{L}_{smooth} = 1 - \cos\left\langle \mathcal{D}_h \times \hat{\mathcal{D}}_w, \mathcal{D}_h^* \times \mathcal{D}_w^* \right\rangle, \quad (9)$$

where D_w and D_h are the gradient vectors of the depth map D along the width and height axes, respectively. When calculating the loss, we discard pixels with depth values falling outside the range [0.3, 1.5] to mitigate the impact of outliers.

4 Experiment

4.1 Dataset

We train and evaluate on three widely used transparent object depth completion datasets: ClearGrasp [6], TransCG [8], and Omniverse [7].

- ClearGrasp Dataset: This is the first large-scale transparent object depth completion dataset, containing both synthetic and real-world scene data. The synthetic training set contains 45,454 RGB-D images generated from CAD models of nine transparent plastic objects. The real test set is divided into two subsets: the real-known subset contains 173 images, including the five transparent objects seen in the training set, and the real-novel subset contains 113 images, including five novel transparent objects not seen in the training set.
- TransCG Dataset: This is a real-world transparent object depth completion dataset containing 57,715 real RGB-D images taken by a robot equipped with an Intel RealSense camera. This dataset covers 51 common transparent/semi-transparent/reflective objects from everyday life, encompassing 130 different real-world scenes. In this paper, we select several subsets of scenarios that do not overlap for experiments in order to reduce data redundancy and better evaluate the generalization ability of the model.
- Omniverse Object Dataset (OOD): This dataset contains 61,317 synthetic RGB-D images, primarily targeting transparent objects in cluttered scenes.

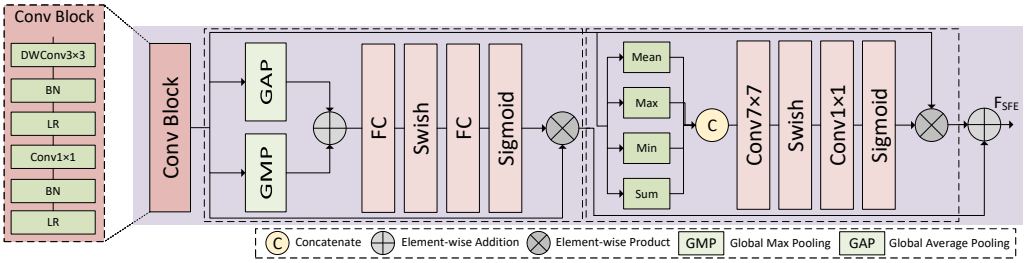


Fig. 4 Structure of the proposed SFE module, which enhances feature representations of structural transitions to facilitate the reconstruction of depth discontinuities

4.2 Evaluation Metrics

We use standard evaluation metrics widely used in the field of transparent object depth completion for evaluation.

- RMSE: Root mean square error between the predicted depth and the ground truth.
- REL: Mean absolute relative error between the predicted depth and the ground truth.
- MAE: Mean absolute error between the predicted depth and the ground truth.
- Threshold precision (δ): Percentage of pixels satisfying $\max(d_i/\hat{d}_i, \hat{d}_i/d_i) < \delta$ for $\delta \in \{1.05, 1.10, 1.25\}$. Higher values indicate better performance.

All evaluation metrics are calculated only in the transparent object areas of the scene, consistent with previous studies.

4.3 Implementation Details

All experiments were conducted on Ubuntu 20.04 using a single NVIDIA GeForce RTX 3090 GPU (24GB VRAM) for both training and testing. During model training, we employed the AdamW optimizer. A multi-stage learning rate scheduling strategy was used, decaying the learning rate by 50% at training epochs 15, 25, and 35. The batch size was set to 16, and the total training epochs were 40. All input images were resized to 320×240 resolution during both training and testing.

To prevent overfitting, we employed extensive data augmentation techniques during training, including random horizontal/vertical flipping, random rotation within ±15°, random adjustments to brightness/contrast/saturation in the HLS color space, and adding Gaussian noise to the original depth images. The weight β of the loss function was set to 0.1 in all experiments.

4.4 Comparison with State-of-the-Art Methods

1) Results on ClearGrasp Dataset

Table 1 presents our comparison results with SOTA methods on the ClearGrasp dataset, including quantitative comparisons on the ClearGrasp Real-known and Real-novel subsets. To ensure a fair comparison, all baseline results were either reproduced using officially released codes with default configurations or directly cited from their original papers. As shown in the table, our TADNet achieves excellent performance on both subsets.

On the ClearGrasp Real-known subset, TADNet achieves RMSE, REL, and MAE values of 0.020, 0.024, and 0.014, comparable to the best-performing method CAGT [13], and outperforms all other methods. In terms of accuracy metrics, TADNet's $\delta_{1.10}$ and $\delta_{1.25}$ are both slightly higher than CAGT, indicating that it has excellent high-precision depth estimation capabilities.

On the ClearGrasp Real-novel subset, TADNet achieves RMSE, REL, and MAE values of 0.013, 0.019, and 0.010, second only to CAGT, but outperforming all other methods. Notably, TADNet achieves a $\delta_{1.10}$ of 99.23%, higher than CAGT's 99.04%, and its $\delta_{1.05}$ and $\delta_{1.25}$ are also the second best methods, only slightly lower than CAGT, which is specifically designed for Sim-to-Real. This demonstrates its robustness to unseen transparent objects. Furthermore, its model size of 12.5MB is smaller than CAGT's 40.7MB. The outstanding generalization performance highlights the effectiveness of the transparency-aware mechanism, which enables mask-guided learning, utilizing masks to represent latent prior information.

In addition, to qualitatively evaluate performance, we present a visualization of the depth completion results in Fig. 5. It can be seen that the original depth map acquired by the RGB-D camera exhibits severe missing values and noise in the transparent object regions. As can be observed, our model effectively repairs the missing depth values in the central region of transparent objects, enabling the output depth map to maintain relatively clear object details and a certain degree of surface smoothness even in cluttered environments.

2) Results on TransCG Dataset

Table 2 presents the quantitative comparison results on the TransCG real-world dataset. Given

Table 1 Depth Completion Results in Comparison on ClearGrasp Dataset

Method	Publications	RMSE↓	REL↓	MAE↓	$\delta_{1.05}$ ↑	$\delta_{1.10}$ ↑	$\delta_{1.25}$ ↑
Train ClearGrasp Synthetic / Test ClearGrasp Real-known							
ClearGrasp [6]	ICRA	0.039	0.055	0.029	72.62	86.96	95.58
LIDF-Refine [7]	CVPR	0.028	0.033	0.020	82.37	92.98	98.63
DepthGrasp [28]	IROS	0.031	0.039	0.021	74.69	89.73	97.35
DFNet [8]	RAL	0.068	0.107	0.059	32.42	56.88	91.47
FDCT [9]	RAL	0.065	0.103	0.057	33.08	59.81	91.70
TODE-Trans [11]	ICRA	0.022	0.029	0.017	84.09	96.25	**99.81**
DistillGrasp [17]	RAL	0.021	0.032	0.018	80.78	94.91	99.56
CAGT [13]	TCSVT	**0.019**	0.024	0.014	**88.74**	97.19	99.74
TADNet (Ours)	-	0.020	**0.024**	**0.014**	88.19	**97.49**	99.75
Train ClearGrasp Synthetic / Test ClearGrasp Real-novel							
ClearGrasp [6]	ICRA	0.034	0.045	0.025	76.72	91.00	97.63
LIDF-Refine [7]	CVPR	0.025	0.036	0.020	76.21	94.01	99.35
DepthGrasp [28]	IROS	0.022	0.033	0.017	82.37	93.46	98.48
DFNet [8]	RAL	0.051	0.088	0.046	31.23	64.66	97.77
FDCT [9]	RAL	0.043	0.073	0.038	39.42	75.54	99.09
TODE-Trans [11]	ICRA	0.014	0.020	0.011	91.63	98.45	99.94
DistillGrasp [17]	RAL	0.020	0.028	0.015	84.12	96.06	99.43
CAGT [13]	TCSVT	**0.012**	**0.017**	**0.009**	**95.23**	99.04	**99.98**
TADNet (Ours)	-	0.013	0.019	0.010	94.40	**99.23**	99.91

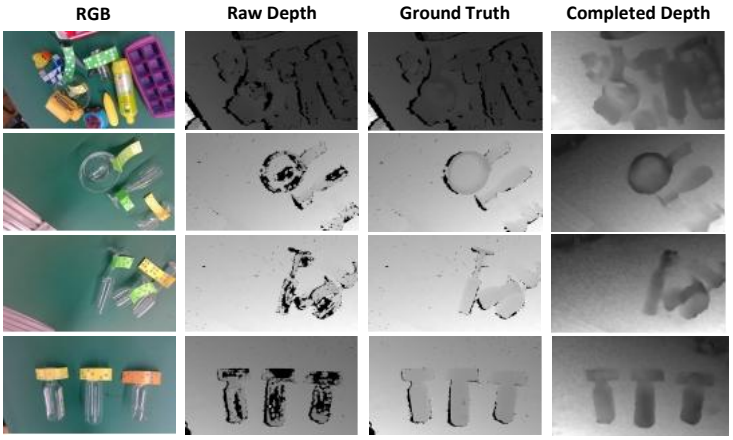


Fig. 5 Visualization of the original depth, the ground truth depth, and the depth completed by TADNet on several testing scenes

the difficulty in obtaining the source code for methods targeting transparent objects, and the significant domain gap in real-world datasets, we evaluate only reproducible methods. The methods are implemented using their official default configurations to ensure the fairness of the comparison. Our analysis focuses on how TADNet addresses the shortcomings of existing methods in real-world scenarios: ClearGrasp and LIDF-Refine perform the worst, possibly because their global optimization and implicit neural representations are sensitive to the distribution of real-world data. While DFNet, TODE-Trans and TDCNet improve performance, their performance still falls short in capturing local details, as indicated by the $\delta_{1.05}$ metric. TADNet achieves the best performance on RMSE, REL, MAE, $\delta_{1.05}$, and $\delta_{1.25}$, benefiting from the joint design of the TAA and SFE modules that balance global reasoning in transparent regions with local structural enhancement.

Table 2 Depth Completion Results in Comparison on TransCG Dataset

Method	Publications	RMSE↓	REL↓	MAE↓	$\delta_{1.05}$ ↑	$\delta_{1.10}$ ↑	$\delta_{1.25}$ ↑
ClearGrasp [6]	ICRA	0.054	0.083	0.037	50.48	68.68	95.28
LIDF-Refine [7]	CVPR	0.019	0.034	0.015	78.22	94.26	99.80
DFNet [8]	RAL	0.024	0.038	0.018	72.54	92.91	99.95
TODE-Trans [11]	ICRA	0.019	0.031	0.014	79.53	96.84	99.98
TDCNet [12]	IEEE Sens. J.	0.017	0.029	0.013	82.66	**98.11**	99.98
TADNet (Ours)	-	**0.017**	**0.028**	**0.013**	**84.21**	97.60	**99.99**

4.5 Cross-Domain Experiment

Although the previous experiments achieved excellent performance, they were all trained and tested on a single dataset. Therefore, cross-domain experiments are needed to verify generalization. Table 3 shows the results of the cross-domain experiment. The model was trained on a combined synthetic dataset of ClearGrasp and OOD and tested on the real-world TransCG dataset.

The poor performance of ClearGrasp may be attributed to its multi-subtask pipeline being highly dependent on the distribution of synthetic data and cannot adapt to changes in the real-world environment. LIDF-Refine even shows a serious performance drop. This may be because its implicit function overfits the synthetic scene and cannot generalize to real-world data. DFNet shows a slight improvement, but due to its limited global context modeling capabilities, it still cannot bridge the gap between synthetic and real-world scenes. Although TranspareNet's δ metric demonstrates significant improvement compared to existing methods, its error metrics perform poorly. This suggests that its strategy of incorporating point clouds aids in enhancing fine-grained accuracy but struggles to address overall depth error issues in cross-domain scenarios. TODE-Trans utilizes the global modeling capabilities of the Transformer to extract generalizable features, but its performance is not optimal due to the lack of task-specific modules. CAGT is designed specifically for simulation-to-real-world transfer and performs best in cross-domain experiments.

TADNet's $\delta_{1.25}$ metric is slightly better than CAGT, and for the remaining metrics, TADNet ranks second after CAGT. This indicates that the method has good simulation to real-world transfer capability and maintains a certain level of depth completion accuracy.

4.6 Ablation Experiment

To verify the impact of these core components in TADNet on network performance, we conducted ablation experiments on the ClearGrasp dataset. All ablation experiments used the same training and testing settings, modifying only the corresponding modules to isolate their contributions. Table 4 presents the results of the ablation experiments, where each core component is removed from the full model to evaluate its individual contribution.

Removing the TAA module increases the RMSE from 0.020 to 0.024; the smaller the δ, the more significantly the accuracy decreases. This indicates that the TAA module plays a crucial role in modeling transparent regions by implementing global contextual reasoning and adaptive feature aggregation.

Replacing the CDCU with ordinary convolutions and upsampling resulted in a decline in all metrics except for $\delta_{1.25}$. This means the dense connections help improve feature reuse and local representation when decoding, which is beneficial for recovering fine geometric structures in depth completion.

Removing the SFE module also resulted in a decrease of all metrics. This performance reduction indicates that the SFE module helps to better represent local structural changes in feature space.

Overall, the full model achieved the best performance across all evaluation metrics. The combination of the TAA, SFE, and CDCU decoder modules results in better depth completion.

5 Conclusion

In this paper, we proposed TADNet, a hierarchical encoder-decoder network that leverages a Swin-Transformer encoder for long-range contextual modeling and a convolutional decoder with CDCU units for dense depth recovery. To further address the difficulties of transparent object sensing, we proposed the TAA module for region-aware feature aggregation and the SFE module to optimize structural representations and reconstruct depth discontinuities. In our experiments on ClearGrasp and TransCG datasets, our proposed method performs competitively in real-world transparent scenes.

Table 3 Depth Completion Results in Comparison on Cross-Domain Experiment

Method	Publications	RMSE↓	REL↓	MAE↓	$\delta_{1.05}$ ↑	$\delta_{1.10}$ ↑	$\delta_{1.25}$ ↑
Train ClearGrasp+OOD / Test TransCG Real							
ClearGrasp [6]	ICRA	0.061	0.108	0.049	33.59	54.73	92.48
LIDF-Refine [7]	CVPR	0.146	0.262	0.115	13.70	26.39	57.95
TranspareNet [10]	CoRL	0.071	0.060	0.036	62.99	82.92	95.93
DFNet [8]	RAL	0.048	0.088	0.039	38.65	64.42	95.28
TODE-Trans [11]	ICRA	0.034	0.057	0.026	64.10	78.86	98.80
CAGT [13]	TCSVT	**0.024**	**0.043**	**0.019**	**70.37**	**90.48**	99.57
TADNet (Ours)	-	0.026	**0.043**	0.020	68.22	88.49	**99.94**

Table 4 Ablation Study on the Effectiveness of Different Components in TADNet

Method	RMSE↓	REL↓	MAE↓	$\delta_{1.05}$ ↑	$\delta_{1.10}$ ↑	$\delta_{1.25}$ ↑
Train ClearGrasp Synthetic / Test ClearGrasp Real-known						
w/o TAA	0.024	0.031	0.018	80.09	94.95	99.72
w/o CDCU	0.023	0.029	0.017	83.06	95.47	99.80
w/o SFE	0.024	0.030	0.018	81.43	94.88	99.58
Full model	**0.020**	**0.024**	**0.014**	**88.19**	**97.49**	**99.75**

Declarations

Funding

This work was supported by Fujian Provincial Natural Science Foundation of China(2026J001056).

Competing interests

The authors have no relevant financial or non-financial interests to disclose.

Data availability

The datasets analysed during the current study are publicly available. The ClearGrasp dataset is available from the corresponding publication [6]. The TransCG dataset is available from its official release [8]. The Omniverse Object Dataset (OOD) is available from the associated study [7].

Author contribution

All authors contributed to the methodology design, algorithm implementation, and experimental evaluation. **Xinyang Cai** performed the conceptualization and writing of the original draft. **Yiwen Qi** and **Hedan Liu** provided critical feedback and revised the manuscript. All authors have read and approved the final version of the manuscript.

References

[1] Fang, H., Wang, C., Gou, M., Lu, C.: GraspNet-1Billion: A large-scale benchmark for general object grasping. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 11444–11453 (2020)

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

[3] Qin, Y., Kumon, M., Furukawa, T.: A data-driven multiple model framework for intention estimation. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), pp. 6458–6464 (2022)

[4] Liu, S., Wang, L., Wang, X.: Multimodal data-driven robot control for human–robot collaborative assembly. Journal of Manufacturing Science and Engineering **144**(5), 051012 (2022)

[5] Sun, T., Zhang, G., Yang, W., Xue, J.-H., Wang, G.: TROSD: A new RGB-D dataset for transparent and reflective object segmentation in practice. IEEE Transactions on Circuits and Systems for Video Technology **33**(10), 5721–5733 (2023)

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

[7] Zhu, L., Mousavian, A., Xiang, Y., Mazhar, H., Eenbergen, J., Debnath, S., Fox, D.: RGB-D local implicit function for depth

completion of transparent objects. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4647–4656 (2021)
[8] Fang, H., Fang, H.-S., Xu, S., Lu, C.: TransCG: A large-scale real-world dataset for transparent object depth completion and a grasping baseline. IEEE Robotics and Automation Letters **7**(3), 7383–7390 (2022)
[9] Li, T., Chen, Z., Liu, H., Wang, C.: FDCT: Fast depth completion for transparent objects. IEEE Robotics and Automation Letters (RAL) **8**(9), 5823–5830 (2023)
[10] Xu, H., Wang, Y.R., Eppel, S., Aspuru-Guzik, A., Shkurti, F., Garg, A.: Seeing glass: Joint point-cloud and depth completion for transparent objects. In: Proceedings of the Conference on Robot Learning (CoRL), pp. 827–838 (2022)
[11] Chen, K., Wang, S., Xia, B., Li, D., Kan, Z., Li, B.: TODE-Trans: Transparent object depth estimation with transformer. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), pp. 4880–4886 (2023)
[12] Fan, X., Ye, C., Deng, A., Wu, X., Pan, M., Luo, S., Yang, H.: TDCNet: Transparent objects depth completion with CNN-transformer dual-branch parallel network. IEEE Sensors Journal (2025)
[13] Jing, X., Qian, K., Vincze, M.: CAGT: Sim-to-real depth completion with interactive embedding aggregation and geometry awareness for transparent objects. IEEE Transactions on Circuits and Systems for Video Technology **35**(7), 6656–6670 (2025)
[14] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30, pp. 5998–6008 (2017)
[15] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 10012–10022 (2021)
[16] Huang, G., Liu, Z., Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4700–4708 (2017)
[17] Huang, Y., Chen, J., Michiels, N., Asim, M., Claesen, L., Liu, W.: Distillgrasp: Integrating features correlation with knowledge distillation for depth completion of transparent objects. IEEE Robotics and Automation Letters **9**(10), 8945–8952 (2024)
[18] Ichnowski, J., Avigal, Y., Kerr, J., Goldberg, K.: Dex-nerf: Using a neural radiance field to grasp transparent objects. In: Proceedings of the Conference on Robot Learning (CoRL), pp. 526–536 (2021)
[19] Duisterhof, B.P., Mao, Y., Teng, S.H., Ichnowski, J.: Residual-nerf: Learning residual nerfs for transparent object manipulation. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), pp. 13918–13924 (2024)
[20] Kim, J., Noh, J., Lee, D.-G., Kim, A.: Transplat: Surface embedding-guided 3d gaussian splatting for transparent object manipulation. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), pp. 3190–3196 (2025)
[21] Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM **65**(1), 99–106 (2021)
[22] Eigen, D., Puhrsch, C., Fergus, R.: Depth map prediction from a single image using a multi-scale deep network. In: Advances in Neural Information Processing Systems, vol. 27, pp. 2366–2374 (2014)
[23] Duan, Y., Guo, X., Zhu, Z.: Diffusiondepth: Diffusion denoising approach for monocular depth estimation. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 432–449 (2024)
[24] Ma, F., Karaman, S.: Sparse-to-dense: Depth prediction from sparse depth samples and a single image. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), pp. 4796–4803 (2018)
[25] Park, J., Joo, K., Hu, Z., Liu, C.-K., Kweon, I.S.: Non-local spatial propagation network for depth completion. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 120–136 (2020)
[26] Zhang, Y., Funkhouser, T.: Deep depth completion of a single RGB-D image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 175–185 (2018)
[27] Cao, Y., Xu, J., Lin, S., Wei, F., Hu, H.: Gcnet: Non-local networks meet squeeze-excitation networks and beyond. In: 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), pp. 1971–1980 (2019)
[28] Tang, Y., Chen, J., Yang, Z., Lin, Z., Li, Q., Liu, W.: DepthGrasp: Depth completion of transparent objects using self-attentive adversarial network with spectral residual for grasping. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 5710–5716 (2021)