ADDFNet: A Robotic Grasping Depth Map Completion Network Integrating Differential Enhancement Convolution and Hybrid Attention

Nan Liu , Yi-Horng Lai * , Yue Wu , Jiaen Wang , Xian Yu

Posted Date: 15 April 2026

doi: 10.20944/preprints202604.1051.v1

Keywords: transparent objects; depth completion; differential enhancement convolution; hybrid attention mechanism; robotic grasping

 doi:10.20944/preprints202604.1051.v1

Article

ADDFNet: A Robotic Grasping Depth Map Completion Network Integrating Differential Enhancement Convolution and Hybrid Attention

Nan Liu, Yi-Horng Lai *, Yue Wu, Jiaen Wang and Xian Yu

School of Mechanical, Electrical Engineering and Automation, Xiamen University Tan Kah Kee College, Zhangzhou 363105, China

* Correspondence: lai81@xujc.com

Abstract

In the field of industrial robot vision, the accurate recognition and localization of transparent objects face multiple challenges. First, depth sensor data suffer from sparsity and non-uniform distribution. Even with high-end LiDAR, the obtained depth maps are generally sparse and severely noisy, especially around object boundaries. Most existing methods assume a fixed sparsity level, leading to significant performance degradation when the actual sparsity varies dynamically. Second, there is a cross-modal feature alignment issue between RGB and depth data. Simple channel stacking or addition neglects the modeling of feature correlations between the two modalities, resulting in insufficient information utilization. Furthermore, existing methods still lack the capability to model the multi-directional gradient variations of transparent objects under complex backgrounds. To address these issues, this paper proposes Attention-based Difference-enhanced Depth Fusion Network (ADDFNet), a depth completion network for transparent objects, which achieves synergistic improvements in accuracy and robustness through two key designs: MDAM and CMFR. To tackle the dynamic variation of sparsity and edge blurring, a Multi-directional Differential Attention Module (MDAM) is designed. It explicitly extracts multi-directional gradient information via multi-branch differential convolutions, enhancing the network's robust perception of sparse edges. Within MDAM, a Detail Enhancement Differential sub-module (DEDM) and a Dynamic Convolution with Symmetry-enhanced Geometry Attention sub-module (DSCA) are introduced to adaptively adjust the focus regions under varying sparsity inputs. To address the insufficient cross-modal feature alignment, a Cross-Modal Feature Refinement (CMFR) module is introduced, which leverages RGB context to progressively guide and enhance depth features at the encoding stage, achieving finer cross-modal feature alignment. Evaluation results on the ClearPose and TransCG datasets demonstrate that ADDFNet outperforms comparison methods in terms of accuracy metrics.

Keywords: transparent objects; depth completion; differential enhancement convolution; hybrid attention mechanism; robotic grasping

1. Introduction

In recent years, with the rapid development of industrial automation and smart manufacturing technologies, robotic grasping tasks have become a core component in fields such as logistics sorting, flexible assembly, and home services [1,2]. In complex operational scenarios, the widespread presence of transparent objects (such as glassware, plastic packaging, and laboratory consumables) poses severe challenges to robot visual perception systems [3,4]. As a key prerequisite for achieving precise grasping, depth information acquired by RGB-D sensors is crucial for computing object poses [5,6]. However, limited by the unique refraction and reflection physical properties of transparent materials, light undergoes path deflection and specular reflection when passing through or contacting object surfaces, causing mainstream depth cameras (such as structured light or time-of-flight cameras) to capture raw

 doi:10.20944/preprints202604.1051.v1

depth maps with large-scale data missing or severe depth drift [7,8]. This failure in depth perception directly restricts the robot's prediction accuracy for grasp poses, easily leading to collisions or grasping failures. Therefore, transparent object depth completion has become a core preprocessing task for enhancing the robustness of robot intelligent operations.

To address the failure of depth perception for transparent objects, early research mainly attempted to use handcrafted geometric priors (such as dark channel prior [9] or color attenuation prior [10]) and multi-view geometric constraints [11,12] to recover depth information. Although these methods have shown certain effectiveness in controlled environments such as laboratories, their generalization capability and robustness face enormous challenges in real industrial scenarios with intense lighting fluctuations or complex background textures, due to excessive reliance on fixed physical assumptions. With the rise of deep learning, multi-stage frameworks represented by ClearPose [13] have opened a new era of end-to-end perception. Such methods typically decompose the completion task into multiple subtasks such as semantic mask prediction, surface normal estimation, and occlusion boundary extraction, and finally reconstruct the depth map through global optimization algorithms (such as Poisson reconstruction or gradient-based refinement). Although this "divide and conquer" strategy significantly improves completion precision, its complex intermediate representation extraction process and high optimization computational cost also significantly increase deployment costs. In recent years, single-stage depth completion networks based on U-Net architecture have gained widespread attention due to their efficient end-to-end learning capability [14,15]. Such networks directly learn the mapping from RGB images to depth maps through encoder-decoder architecture, demonstrating excellent inference performance [16]. However, these methods still face visual challenges unique to transparent objects: on one hand, the limited receptive field of conventional convolution operators makes it difficult to effectively capture global spatial context of large transparent objects, resulting in local structural distortion in completed depth maps; on the other hand, the smoothing effect of standard convolution tends to blur the already sparse edge features of transparent objects, leading to loss of geometric details.

In existing depth completion research, how to further improve structural consistency and robustness under high-precision prediction remains the main challenge. On one hand, visual edges of transparent objects are typically extremely sparse and susceptible to background noise interference, making traditional convolution operators unable to explicitly capture multi-directional gradient variations, resulting in lack of geometric consistency in completed object boundaries; on the other hand, depth features in transparent regions have high spatial heterogeneity, and fixed convolution kernel weights cannot adaptively adjust attention regions according to input scenes, limiting the network's modeling capability for complex transparent structures. Furthermore, how to efficiently fuse multi-modal inputs and multi-scale features while avoiding feature degradation during depth propagation is also a key issue that needs to be addressed.

Based on the above challenges, this paper proposes a transparent object depth completion network named ADDFNet, aiming to improve completion accuracy and structural recovery stability through refined feature enhancement and adaptive attention mechanisms. Specifically, we first design MDAM, which explicitly extracts spatial gradients through integrated multi-branch differential convolution operators, and introduces two sub-components within MDAM: DEDM and DSCA, using dynamically generated convolution kernels to achieve adaptive adjustment of features, combined with spatial symmetry optimization for interactive modeling, enabling the network to accurately focus on key depth features in transparent regions; we then introduce CMFR during the encoding stage, using RGB context to guide and enhance depth features layer by layer at each downsampling level, further improving model robustness.

The main contributions of this paper are summarized as follows:

- Proposed a novel transparent object depth completion network ADDFNet, which effectively addresses the problems of blurred edge perception and insufficient feature modeling capability for transparent objects through deep integration of multi-directional differential enhancement and hybrid attention mechanisms.

 doi:10.20944/preprints202604.1051.v1

- Designed MDAM, which utilizes differential convolution to explicitly enhance gradient feature extraction, greatly improving the restoration accuracy of transparent object contours in complex backgrounds.
- Constructed DEDM and DSCA within MDAM, and designed CMFR, achieving adaptive feature perception and efficient integration of multi-source information, improving the accuracy and stability of depth prediction.
- ADDFNet was evaluated on two public datasets: ClearPose and TransCG [17], with experimental results verifying the effectiveness of the proposed method in transparent object depth completion tasks.

2. Related Work

2.1. General Depth Completion Methods

Early studies on general depth completion primarily focused on fusing sparse depth measurements with RGB cues. Representative work such as Sparse-to-Dense [18] validated the effectiveness of end-to-end regression in outdoor sparse LiDAR scenarios by jointly encoding color texture and sparse observations. Deep Depth Completion [19] further highlighted the importance of jointly modeling local geometric structures and global context, thereby improving completion quality in complex structural regions. Building upon these advances, CSPN [20] introduced a context-aware spatial propagation network that progressively fills missing depth values through iterative neighborhood propagation. GuideNet [21] proposed a guided upsampling architecture, leveraging multi-scale feature fusion to better recover fine-grained structures. Meanwhile, FuseNet [22] designed a multimodal fusion module to effectively integrate RGB semantic information with geometric cues from depth maps, laying an important foundation for cross-modal feature interaction.

As the field progressed, subsequent methods increasingly emphasized structural propagation and neighborhood consistency constraints. NLSPN [23] significantly improved depth continuity in edge and thin-structure regions by adaptively learning pixel-wise correlations through non-local spatial propagation. At the same time, ACMNet [24] proposed an adaptive context-aware multi-scale architecture, enhancing modeling capacity across different structural scales via dynamic receptive-field adjustment, while PENet [25] adopted a progressive enhancement strategy to iteratively refine depth predictions in a coarse-to-fine manner. For transparent-scene depth completion, researchers further incorporated physical priors and local implicit modeling into general frameworks. For example, RGB-D Local Implicit Function [26] improved surface reconstruction in transparent regions through local implicit representations. In addition, DeepLiDAR [27] and Penet [28] provided new technical perspectives on sparse-to-dense depth conversion from the viewpoints of data synthesis and penetration-aware completion, respectively. Inspired by these studies, our method draws on multi-modal fusion and cross-modal feature interaction, and integrates adaptive context-aware modeling with progressive enhancement to improve sparse-edge perception and depth completion accuracy for transparent objects.

2.2. Transparent Object Depth Completion

Due to the unique refraction and reflection physical properties of transparent objects, mainstream depth sensors often produce severe depth missing or geometric distortion during acquisition due to light path deflection. Early research mainly relied on multi-stage optimization frameworks, among which Chen et al. [13] proposed ClearPose as a milestone. This method decouples depth completion into multiple subtasks such as surface normal estimation, mask prediction, and boundary extraction, and reconstructs the depth map using global optimization algorithms, achieving significant breakthroughs in accuracy. However, the complex inference process and substantial computational cost limit its application in large-scale deployment scenarios.

To address the inefficiency of multi-stage models, single-stage end-to-end networks have gradually become the mainstream direction in this field. Compared with traditional methods, deep neural

 doi:10.20944/preprints202604.1051.v1

networks can more effectively capture cross-modal correlations through hierarchical feature learning. Seeing Glass [14] innovatively introduced Haar Wavelet Transform (HWT) to construct a cascaded framework to address high-frequency detail loss, achieving lightweight while effectively preserving geometric edges in transparent regions. DistillGrasp [15] utilized knowledge distillation technology to transfer long-range spatial correlations captured by Transformer teacher networks to efficient CNN student networks, improving completion quality and structural consistency through Position Correlation Blocks (PCB). Additionally, TCRNet [16] proposed a cascaded refinement architecture, combining Transformer error modules to calibrate global depth distribution.

Subsequent research has focused on breaking through the limitations of single modality by mining multi-source cues. DualTransNet [29] no longer relies solely on binary segmentation masks, but instead guides internal structure recovery by mining rich semantic features from intermediate layers of segmentation networks. SRNet-Trans [30] and Voxel-DL [31] respectively strengthened spatial constraints from the perspectives of geometric convolution and 3D voxel space. Although existing methods have made considerable progress in feature enhancement and adaptive perception, achieving robust geometric perception in extremely sparse edge regions severely affected by background noise remains a weak point. These explorations provide important theoretical support for the multi-directional differential enhancement mechanism proposed in this paper.

2.3. Robotic Grasp Detection

Mainstream robotic grasp detection is divided into 2D rectangle-based planar grasping [2] and six-degree-of-freedom (6-DOF) spatial grasping [3], with performance highly dependent on the completeness of input geometric information. However, the unique refraction and reflection characteristics of transparent objects often cause depth maps to produce severe missing and geometric drift [7], directly restricting the prediction accuracy of grasp poses. To address this challenge, works such as ClearPose [13] and TCRNet [16] have improved perception robustness through multi-stage optimization or cascaded refinement. Although these methods have made progress in feature capture, they still struggle to effectively model multi-directional gradient variations in extremely sparse edge regions severely affected by background noise, leading to blurred object contours or loss of local details. Therefore, this paper integrates ADDFNet, utilizing multi-directional differential enhancement to explicitly strengthen edge perception, combined with a hybrid attention mechanism based on geometric symmetry enhancement to achieve adaptive adjustment of features. Experiments show that this solution significantly reduces information loss and provides precise and geometrically consistent depth constraints for downstream grasping tasks in complex scenes.

3. Method

In this section, we first elaborate on the hierarchical design philosophy of the overall ADDFNet network architecture. Subsequently, we provide technical implementation details for two core modules: MDAM (containing DEDM and DSCA sub-components) and CMFR. These core designs collectively construct a transparent object depth completion framework for complex environments through explicit edge gradient enhancement, adaptive feature modeling, and multi-source information collaborative refinement.

3.1. Overview

This paper designs a new transparent object depth completion network ADDFNet, whose overall architecture is shown in Figure 1. The network adopts an encoder-decoder structure, taking RGB images and sparse depth maps as dual-modal inputs.

In the encoder part, the network first downsamples the preliminarily fused RGB and depth features through a convolutional layer. Subsequently, deep features are gradually extracted through four cascaded modules, each centered around MDAM for feature extraction. To effectively fuse multi-modal information, the network uses depth maps of corresponding scales as auxiliary inputs at each level of the encoder, and designs a cross-modal channel attention mechanism CMFR, which

 doi:10.20944/preprints202604.1051.v1

uses features from the main branch to generate adaptive channel weights for depth features, thereby guiding the network to focus on key geometric structures. Additionally, the encoder has parallel connection branches in the first three downsampling stages, whose outputs are passed to the decoder as skip connections.

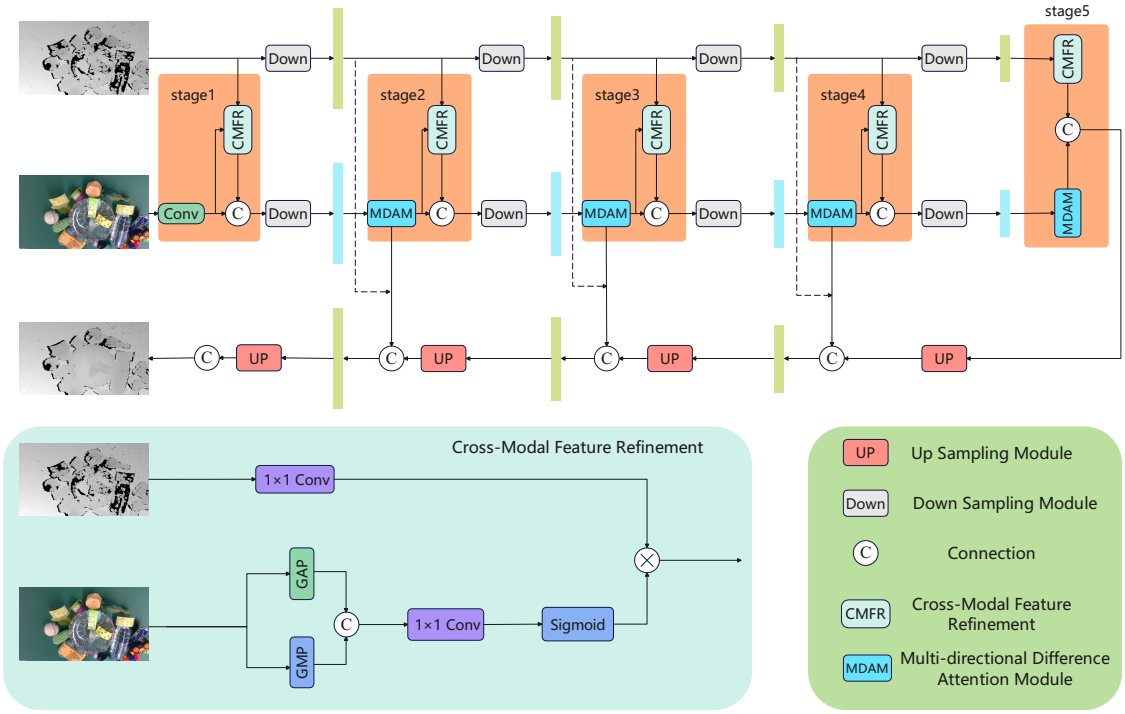

Figure 1. Overall network architecture of ADDFNet, where the MDAM module consists of two sub-modules: DEDM and DSCA.

In the decoder part, the network gradually restores the spatial resolution of feature maps through four symmetric decoding modules. Starting from the second decoding module, each module's input consists of upsampled features from the previous level concatenated with skip connection features from the corresponding encoder level. Within each decoding module, the feature flow first passes through convolutional layers and MDAM for fusion and refinement, and finally resolution is increased through efficient upsampling convolution. Ultimately, the network generates a dense completed depth map with the same resolution as the input RGB image through an output module.

3.2. CMFR

In existing depth completion methods, cross-modal feature fusion typically employs simple concatenation or addition operations [13,17], or uses 3D spatial constraints and geometric convolution to enhance spatial consistency [30,31]. Such fusion is often performed at the end of the encoder or during decoding [15,29], failing to fully utilize the guiding role of RGB images on depth features during feature extraction. Inspired by the aforementioned cascaded refinement structure ideas [14,16], we believe that feature refinement should not be limited to decoding or post-processing stages, but should repeatedly use RGB information to gradually enhance depth features during the encoding process. Based on this idea, we design CMFR in ADDFNet, as shown in the blue region in Fig. 1, aiming to selectively enhance depth feature representations using contextual information provided by RGB images. Unlike existing methods that place refinement in the decoding stage, the CMFR module is embedded into each downsampling level of the encoder to guide the network to focus on geometric structures crucial for depth completion during the feature extraction stage. We believe that introducing RGB-guided mechanisms at each level can more effectively suppress noise interference and improve the representation quality of depth features.

 doi:10.20944/preprints202604.1051.v1

Specifically, for the i-th encoder layer, we first perform preliminary fusion of RGB features and depth features:

$$F^i_{RGB} \in \mathbb{R}^{B\times C\times H\times W}, \quad F^i_D \in \mathbb{R}^{B\times 1\times H\times W}, \tag{1}$$

where B, C, H, and W represent batch size, number of channels, feature map height, and width, respectively. Considering that RGB features contain richer semantic information, we use them to guide and enhance depth features. The implementation of the CMFR module mainly includes three steps.

First, in the channel attention generation stage, we perform Global Average Pooling (GAP) and Global Max Pooling (GMP) on RGB features, concatenate them, and then use 1×1 convolution to compress them into channel descriptors to capture global responses of each channel:

$$F^i_{GAP} \in \mathbb{R}^{B\times C\times 1\times 1}, \tag{2}$$

Subsequently, in the adaptive weight calculation stage, we model inter-channel dependencies through a bottleneck structure consisting of two fully connected layers (FC), and generate adaptive channel weights $W^i \in \mathbb{R}^{B\times C\times 1\times 1}$ using the Sigmoid function:

$$W^i = \text{Sigmoid}(\text{FC}_2(\text{ReLU}(\text{FC}_1(\text{Concat}(F^i_{\text{GAP}}, F^i_{\text{GMP}}))))) \tag{3}$$

where FC_1 and FC_2 represent the first and second fully connected layers, respectively, and ReLU is the activation function. Finally, in the depth feature weighting stage, we apply the learned channel weights to depth features to achieve adaptive enhancement of depth features:

$$\hat{F}^i_D = W^i \cdot F^i_D \tag{4}$$

where $\hat{F}^i_D$ represents the enhanced depth features, and $\cdot$ denotes element-wise multiplication along the channel dimension. Through this approach, the network can selectively enhance channel features important for the depth completion task while suppressing irrelevant information, thereby improving the discriminative capability of depth features. After this, the enhanced depth features are concatenated with RGB features and fed into MDAM to extract more robust edge structure information.

3.3. MDAM

3.3.1. Overall Design

To address the problem of extremely sparse edge features of transparent objects that are susceptible to background noise, light refraction, and reflection interference, existing research mainly relies on traditional convolutional neural networks (CNN) for local spatial feature aggregation. However, standard convolution kernels are essentially weighted sums of features within local neighborhoods, which tend to produce feature smoothing effects when extracting high-frequency gradient information, causing fine contour information of transparent objects to gradually be submerged during layer-by-layer propagation. Additionally, existing attention mechanisms such as SENet [32] and CBAM [33], while able to re-weight features in channel or spatial dimensions, still compute weights based on global or local intensity distributions, lacking explicit modeling of local pixel gradient variations, making it difficult to capture multi-directional subtle edge differences. Therefore, transparent object boundary recognition still faces significant challenges in industrial scenes with intense lighting changes and complex background textures.

Inspired by the differential computation idea in traditional edge detection operators, we note that operators such as Sobel and Canny precisely locate edges by computing intensity changes within local neighborhoods, and this explicit gradient modeling approach is extremely sensitive to edge information. In recent years, differential convolution networks (such as pixel difference networks [34,35] and center differential convolution [36]) have further introduced this idea into learnable network structures, effectively improving network capability for edge and texture detail extraction by introducing pixel-level difference operations during convolution. These studies collectively indicate

 doi:10.20944/preprints202604.1051.v1

that explicitly computing multi-directional pixel-level differences within local neighborhoods can greatly enhance responses to blurred edges. However, existing differential convolution methods mostly focus on general edge or texture enhancement and have not yet fully combined feature projection and attention mechanisms for targeted optimization. If multi-directional differential computation can be combined with feature projection mapping and adaptive weighting mechanisms, the network can strengthen geometric information highly correlated with transparent region boundaries while preserving global context.

Based on these observations, we design MDAM, which strengthens the network's perception of fine transparent object contours through explicit modeling of spatial gradient variations. The module first computes local pixel differences along multiple directions to extract fine-grained edge responses. It then aligns differential responses with original features via feature projection mapping and introduces adaptive hybrid attention [33,37], enabling dynamic focus on the most discriminative boundary regions. After multi-directional fusion, MDAM further applies an attention enhancement unit to improve projection alignment and global contextual awareness. Architecturally, MDAM avoids complex recurrent or deeply stacked structures and adopts a lightweight single-layer computation paradigm for explicit edge enhancement.

Specifically, given the fused feature containing RGB and depth information $\mathbf{X}_{in} \in \mathbb{R}^{(C_{in}+C_{depth})\times H\times W}$, MDAM first projects it into a unified feature space through a 1×1 convolution, followed by batch normalization ($\mathcal{B}$) and ReLU activation for cross-channel integration and nonlinear alignment:

$$\mathbf{X}_{proj} = \mathrm{ReLU}\big(\mathcal{B}(\mathrm{Conv}_{1\times1}(\mathbf{X}_{in}))\big), \quad \mathbf{X}_{proj} \in \mathbb{R}^{C\times H\times W} \tag{5}$$

where C denotes the projected channel dimension.

After projection, MDAM enhances $\mathbf{X}_{proj}$ progressively using two core components: DEDM and the DSCA unit. DEDM focuses on multi-directional local gradients and explicitly strengthens fine-edge perception, while DSCA introduces global geometric constraints by combining the stability of static large-kernel convolution and the adaptivity of dynamic convolution to improve long-range spatial consistency. The two components are applied sequentially to achieve comprehensive modeling of transparent object geometry.

3.3.2. DEDM

In transparent object depth completion, standard convolution layers lack explicit constraints on local pixel-wise gradient variation and therefore tend to overemphasize intensity-level cues while weakening high-frequency edge and texture responses. This issue is especially severe in transparent scenes, where boundaries between transparent regions and backgrounds often have very low contrast and are strongly affected by refraction, reflection, and illumination changes. To introduce explicit gradient priors into the network, we introduce DEDM, shown in Figure 2, as a core component inside MDAM.

Existing studies have shown that differential convolution can improve representation capacity and generalization by convolving pixel-level differences. Typical forms include center differential convolution (CDC) and angular differential convolution (ADC). To better fit transparent object boundaries that are sparse and direction-sensitive, we further introduce horizontal differential convolution (HDC) and vertical differential convolution (VDC), and integrate priors from Sobel [38], pixel difference convolution [35], and large-kernel design [39]. By deploying CDC, HDC, VDC, ADC, and standard convolution in parallel, the network jointly captures intensity-level and multi-directional gradient-level cues for more complete structure modeling.

 doi:10.20944/preprints202604.1051.v1

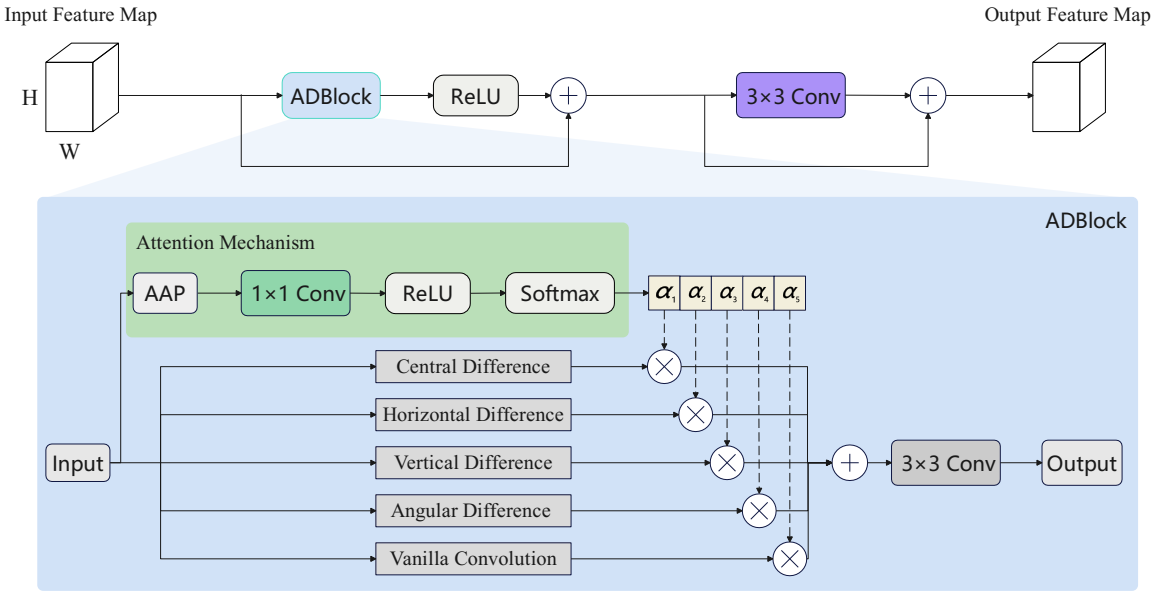

Figure 2. DEDM.

In implementation, DEDM adopts a five-branch parallel structure corresponding to CDC, HDC, VDC, ADC, and standard convolution. To preserve dynamic adaptability while introducing directional gradient priors, we design an adaptive fusion mechanism. For projected feature $\mathbf{X}_{proj}$, adaptive branch weights $\boldsymbol{\alpha}$ are generated by global average pooling (GAP) and two 1×1 convolutions:

$$\boldsymbol{\alpha} = \mathrm{Softmax}\big(\mathrm{MLP}(\mathrm{GAP}(\mathbf{X}_{proj}))\big), \quad \boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_5] \tag{6}$$

The enhanced differential feature is obtained by weighted fusion of all branches:

$$\mathbf{x}_{diff} = \sum_{i=1}^{5} \alpha_i \cdot \Phi_i(\mathbf{X}_{proj}) \tag{7}$$

where $\Phi_1 \sim \Phi_4$ denote CDC, HDC, VDC, and ADC, respectively, and Φ_5 is standard spatial convolution for preserving base semantic cues. This design allows the network to adaptively balance intensity and directional gradient cues under different inputs, improving robust edge enhancement in complex backgrounds.

3.3.3. DSCA

After DEDM extracts and fuses multi-directional gradients, the network obtains enhanced feature $\mathbf{x}_{diff}$ that is sensitive to boundary changes. However, transparent object completion also requires long-range geometric consistency in large smooth regions, where local differential operations alone may be insufficient and can lead to structural discontinuity. To address this, we design a DSCA unit, which combines static large-kernel convolution and dynamically generated convolution kernels for complementary global-context modeling and local adaptive enhancement. In the design of convolutional attention mechanisms, existing research typically employs fixed-weight large-kernel convolutions to expand the receptive field [40,41], and further enhances long-range modeling capability by increasing convolution kernel size [42]; or utilizes dynamic convolutions to adaptively adjust convolution kernel parameters based on input content [37,43]. The former can provide stable long-range spatial interactions, but its weights remain unchanged once trained, making it difficult to handle the complex and variable depth distributions of transparent objects in different scenarios; the latter has strong adaptability but relies entirely on input-generated convolution kernels, which can easily produce unstable gradient updates during early training stages and has relatively limited capability for modeling global structures. Inspired by structural reparameterization techniques [44,45]

 doi:10.20944/preprints202604.1051.v1

and geometric symmetry enhancement concepts [46], we believe that organically integrating the stability of static large kernels with the adaptability of dynamic convolutions is an effective approach to enhance feature representation capability.

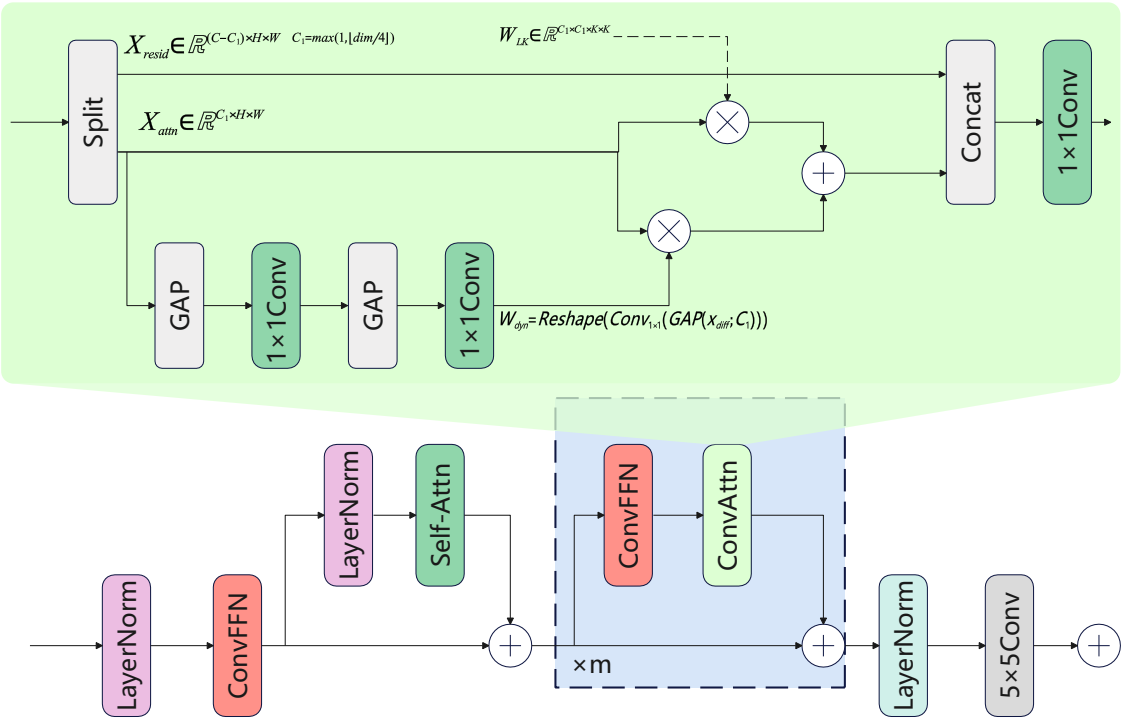

Figure 3. DSCA unit.

Given $\mathbf{x}_{diff} \in \mathbb{R}^{C\times H\times W}$, DSCA first splits channels into an attention branch and a residual branch:

$$\mathbf{x}_{attn} \in \mathbb{R}^{C_1\times H\times W}, \quad \mathbf{x}_{resid} \in \mathbb{R}^{(C-C_1)\times H\times W}, \tag{8}$$

where $C_1 = C//4$ is the number of channels in the attention branch. This design alleviates feature degradation during enhancement by preserving partial original channel information on one hand, and focuses the attention mechanism on the most discriminative channel subset on the other hand, reducing computational overhead. For the attention branch $\mathbf{x}_{attn}$, DSCA deploys two types of convolution operations in parallel: one is static large-kernel convolution, using shared convolution kernels processed with geometric symmetry enhancement (Geo-ensemble):

$$\mathbf{W}_{LK} \in \mathbb{R}^{C_1\times C_1\times K\times K} \quad (\text{taking} K = 13), \tag{9}$$

Through eight-direction geometric transformation averaging, the convolution kernel produces balanced responses to edges and structures in all directions, thereby providing stable long-range spatial interaction; the other is dynamic depth convolution [37,43], which generates dynamic convolution kernels $\mathbf{W}_{dyn}$ from input features through a lightweight network, achieving adaptive adjustment of local details. The dynamic kernel generation process can be expressed as:

$$\mathbf{W}_{dyn} = \text{Reshape}\Big(\text{Conv}_{1\times 1}\Big(\text{GAP}\Big(\mathbf{x}_{diff}[:, :C_1]\Big)\Big)\Big), \tag{10}$$

where GAP represents global average pooling and $\text{Conv}_{1\times 1}$ is a pointwise convolution layer. The dynamic kernel has an explicit size of $\mathbf{W}_{dyn} \in \mathbb{R}^{C_1\times 1\times K_{dyn}\times K_{dyn}}$, where $K_{dyn} = 5$ denotes the spatial kernel size. The dynamic depth convolution uses grouped convolution with the group number set to $G = C_1$, so each channel independently applies one dynamically generated kernel, achieving channel-wise adaptive modeling while maintaining parameter efficiency. Notably, the input channel number

 doi:10.20944/preprints202604.1051.v1

used for dynamic kernel generation is set as $C_{\text{proj}} = C_1$, which guarantees consistency between channel splitting and kernel generation: when the input channel dimension C changes, $C_1 = C//4$ is adjusted accordingly, and both the dynamic kernel size and the group configuration scale consistently with this split. Through this design, dynamic kernels adaptively adjust local response weights according to input-dependent spatial distributions, thereby selectively enhancing transparent object boundaries and internal structures. Adding the outputs of static large-kernel convolution and dynamic depth convolution yields the enhanced result of the attention branch:

$$\mathbf{x}'_{attn} = \text{Conv}_{LK}(\mathbf{x}_{attn}, \mathbf{W}_{LK}) + \text{DDC}(\mathbf{x}_{attn}, \mathbf{W}_{dyn}), \tag{11}$$

where Conv_{LK} represents the static large-kernel convolution operation, and DDC represents the dynamic depth convolution operation. Finally, the enhanced attention branch is concatenated with the original residual branch along the channel dimension to obtain the output features of DSCA:

$$\mathbf{X}_{out} \in \mathbb{R}^{C\times H\times W}. \tag{12}$$

Throughout the computation process, the DSCA unit avoids complex loop or stacking structures, adopting a single-layer parallel computation paradigm, maintaining stable feature enhancement capability while introducing global geometric constraints.

3.4. Loss Function

In transparent object depth completion tasks, the design of the loss function directly determines the network's balancing capability between numerical accuracy and geometric consistency. Single pixel-level loss functions (such as mean squared error), while ensuring predicted depth values approximate ground truth numerically, struggle to effectively constrain structural integrity at transparent object edges; relying solely on perceptual loss or geometric constraints may cause global scale drift. Therefore, inspired by multi-task optimization ideas, this paper constructs a composite loss function consisting of mean squared error loss, perceptual contrast loss, and surface normal smoothness loss, jointly supervising the network from three dimensions: pixel-level accuracy, feature-level structure, and local geometric continuity, guiding ADDFNet to strengthen geometric modeling capability for transparent object edges and surfaces while maintaining depth numerical accuracy. The total loss function is defined as follows:

$$L_{total} = L_{MSE} + L_{contrast} + \beta L_{smooth} \tag{13}$$

where β is the weight coefficient balancing surface smoothness. The specific designs of each loss term are as follows:

Masked Mean Squared Error Loss L_{MSE}: This loss term constrains numerical consistency between predicted depth D_{pred} and ground truth D_{gt} from a pixel-level perspective. Considering that transparent object depth datasets contain large invalid depth regions caused by refraction and reflection, we compute loss only on valid pixels covered by mask Ω, avoiding interference from invalid regions on gradient updates:

$$L_{MSE} = \frac{1}{|\Omega|}\sum_{i\in\Omega}(D_{pred,i} - D_{gt,i})^2 \tag{14}$$

Perceptual Contrast Loss $L_{contrast}$: To address the problem of sparse edge features in transparent objects where conventional loss functions easily lead to boundary blurring, we introduce perceptual contrast loss [47], enhancing the network's modeling capability for high-frequency details through distance constraints in feature space. Specifically, we use the VGG19 network [48] pretrained on ImageNet [49] to extract multi-scale features from predicted depth maps and ground truth depth maps, and compute $L1$ distances between features at each layer:

$$L_{contrast} = \sum_j w_j \|\phi_j(D_{pred}) - \phi_j(D_{gt})\|_1 \tag{15}$$

 doi:10.20944/preprints202604.1051.v1

where ϕ_j represents the j-th layer feature map of VGG19 [48], and w_j is the contribution weight of the corresponding layer. Through this loss term, the network can perceive structural differences between predicted results and real depth in high-dimensional feature space, thereby recovering fine edge contours while maintaining global layout.

Surface Normal Smoothness Loss L_{smooth}: Since robotic grasping tasks are highly sensitive to geometric consistency of object surfaces, relying solely on pixel-level and feature-level losses cannot guarantee local surface smoothness. Therefore, we further introduce surface normal smoothness loss, improving local surface geometric quality by constraining angular differences between surface normals derived from predicted depth maps and ground truth depth maps. This loss is defined as:

$$L_{smooth} = \frac{1}{|\Omega'|} \sum_{i \in \Omega'} (1 - \cos\langle \mathbf{n}_{pred,i}, \mathbf{n}_{gt,i} \rangle) \tag{16}$$

where $\mathbf{n}$ represents surface normal vectors derived from depth maps through finite difference methods, and Ω' is the corresponding expanded mask (eroding the original mask to eliminate edge effects in normal computation at boundaries). By minimizing cosine distance between normal vectors, this loss can effectively suppress local depth mutations, guiding the network to generate smoother depth maps that conform to the true geometric shape of objects.

The above three losses form complementary relationships from three dimensions: numerical accuracy, structural integrity, and local smoothness, enabling ADDFNet to balance depth completion accuracy and geometric fidelity in complex scenes, providing reliable geometric priors for grasp pose estimation.

4. Experiments

4.1. Datasets

To evaluate the performance of ADDFNet in transparent object depth completion tasks, this paper selects two public datasets, ClearPose and TransCG, for training and testing. These two datasets represent two typical data composition types: "synthetic-real hybrid" and "large-scale real scenes".

The ClearPose dataset was first proposed by Chen et al. [13] in 2022. As a pioneering benchmark in the field of transparent object depth perception, this dataset adopts a construction strategy combining synthetic data and real-world data. In the synthetic data part, the dataset contains nine transparent plastic objects with different geometric forms, four of which are specifically used for testing procedures. The synthetic training set contains 18,000 images, the synthetic validation set contains 500 images, and the synthetic test set consists of 400 images, forming a relatively complete train-validation-test split. In contrast, the real-world data in this dataset is relatively limited, mainly used to evaluate the generalization performance of networks from synthetic to real domains. The real-world validation dataset contains five objects from the synthetic dataset, totaling 173 images; the real-world test set introduces five new transparent objects not present in the training set, containing 113 images captured by cameras in real scenes. The ClearPose dataset not only provides RGB images and corresponding sparse depth maps but also annotates scene segmentation labels, camera and object poses, aligned depth, and surface normal information, providing rich supervision signals for multi-task learning and geometric constraint optimization.

The TransCG dataset was proposed by Fang et al. [17] in 2022 and is currently the largest real-world transparent object depth completion dataset. This dataset contains 57,715 pairs of RGB images and corresponding depth maps, covering 51 transparent objects and approximately 200 opaque objects. All images are collected from different real scenes, forming about 130 different scene configurations. Objects in the dataset are randomly distributed in simple or complex scenes, simulating real-world robotic grasping scenarios, with high scene diversity and challenge. To maintain consistency with the original dataset split, we adopt the same data partitioning scheme: the training set uses 34,191 images, and the test set uses 23,524 images. By constructing a large-scale real-scene image collection specifically for transparent objects, the TransCG dataset provides valuable resources for researchers and

 doi:10.20944/preprints202604.1051.v1

practitioners to train and evaluate models tailored to the unique challenges of transparent objects. This dataset not only meets the requirements of professional datasets for considering complex characteristics of transparent objects but also promotes the development of robotic grasping technology through rich scene configurations and object combinations.

From the perspective of data characteristics comparison, the ClearPose dataset provides precise geometric annotations and controllable noise environments in its synthetic data part, facilitating network learning of basic physical properties of transparent objects; while the TransCG dataset, through large-scale real-scene collection, is closer to complex lighting conditions and background interference in actual industrial applications. This complementarity enables joint evaluation on both datasets to reflect algorithm performance in different application scenarios. In the experiments of this paper, we train and test ADDFNet on these two datasets separately and analyze its generalization performance in synthetic and real scenes, providing reference basis for practical deployment of transparent object depth completion technology.

4.2. Experimental Details

All experiments in this paper are conducted in the Ubuntu 24.04 operating system environment, using a single NVIDIA GeForce RTX 5090D graphics card for network training and inference. The deep learning framework uses PyTorch 2.8.0, with CUDA 12.8 acceleration for training. Network training employs the Adam optimizer with an initial learning rate of 10^{-3}. To improve training stability and accelerate convergence, we use a multi-step learning rate scheduler, decaying the learning rate by a factor of 0.2 after the 5th, 15th, 25th, and 35th epochs. The entire training process consists of 40 epochs, with each epoch taking approximately 1 hour. Batch size is set to 16 to balance GPU memory limitations and training stability.

In the data preprocessing stage, we follow the original settings of the TransCG [17] and ClearPose [13] datasets. For the TransCG dataset, input image resolution is uniformly adjusted to 320x240 pixels, with depth values normalized to the [0.3, 1.0] meter range. During training, data augmentation strategies such as random horizontal flipping, color jittering, and brightness adjustment are employed, with an augmentation probability of 0.8 to improve model robustness to lighting changes and viewpoint differences. During testing, all data augmentation operations are disabled to ensure objective evaluation results. Processing of the ClearPose dataset follows its official recommendations, maintaining the split scheme between synthetic and real-world data.

In terms of loss function configuration, we use masked mean squared error loss L_{MSE} as the basic supervision signal, computing loss only on valid depth pixels to avoid interference from invalid regions on gradient updates. To further improve the recovery quality of edge details, we introduce perceptual contrast loss $L_{contrast}$, using the pretrained VGG19 network to extract multi-scale features and constrain structural similarity between predicted depth and real depth in feature space. Simultaneously, to enhance geometric smoothness of local surfaces, we introduce surface normal smoothness loss L_{smooth}, suppressing local depth mutations by constraining angular differences between surface normals derived from predicted depth maps and ground truth depth maps. The total loss function is defined as $L_{total} = L_{MSE} + L_{contrast} + \beta L_{smooth}$, where β is the weight coefficient balancing surface smoothness, set to 0.001 in experiments.

In terms of evaluation metrics, we adopt standard metrics widely used in the field of transparent object depth completion, including Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Relative Error (REL), and accuracy at different thresholds (Threshold@1.05, Threshold@1.10, Threshold@1.25). All metrics are computed within valid depth pixel regions (mask-covered regions) to ensure fairness and comparability of evaluation results. For the ClearPose dataset, we evaluate on the real-world test set. For the TransCG dataset, we follow its official partitioning scheme, evaluating performance on 23,524 test images.

 doi:10.20944/preprints202604.1051.v1

4.3. Evaluation Metrics

To comprehensively evaluate the performance of ADDFNet in transparent object depth completion tasks, this paper follows the standard evaluation system in the field of transparent object depth perception [50,51], adopting multiple quantitative metrics to measure differences between predicted depth maps and ground truth depth maps from different dimensions. Specifically, this paper uses the following four core metrics:

1) RMSE: RMSE is the square root of the average of squared errors, consistent with the original depth unit, facilitating intuitive comparison of absolute error levels between different methods. Its calculation formula is as follows:

$$RMSE = \sqrt{\frac{1}{|D|}\sum_{d\in D}|d-d^*|^2} \tag{17}$$

2) MAE: MAE directly computes the average of absolute differences between predicted depth and ground truth depth, intuitively reflecting the absolute magnitude of prediction errors. Its calculation formula is as follows:

$$MAE = \frac{1}{|D|}\sum_{d\in D}|d-d^*| \tag{18}$$

3) REL: REL measures the relative difference between predicted depth and ground truth depth, reflecting the accuracy of prediction results on a relative scale. Its calculation formula is:

$$REL = \frac{1}{|D|}\sum_{d\in D}\frac{|d-d^*|}{d^*} \tag{19}$$

4) Threshold Accuracy: This metric computes the proportion of pixels where the maximum ratio between predicted depth and ground truth depth is less than a given threshold δ, reflecting the accuracy of prediction results within a specific tolerance range. This paper uses three common thresholds $\delta = 1.05, 1.10, 1.25$ for evaluation, with the judgment condition being:

$$\max\left(\frac{d}{d^*}, \frac{d^*}{d}\right) < \delta \tag{20}$$

where $\delta_{1.05}$, $\delta_{1.10}$, and $\delta_{1.25}$ represent the proportion of pixels with prediction errors within 5%, 10%, and 25% ranges, respectively.

Considering that transparent object depth datasets contain large invalid depth regions caused by refraction and reflection, this paper uses valid pixel masks in the computation of all the above metrics, only counting results within mask-covered regions to ensure fairness and comparability of evaluation.

The above metrics comprehensively evaluate the performance of depth completion algorithms from different dimensions: RMSE emphasizes overall deviation at the squared error level, MAE reflects absolute error level, REL reflects relative error, and threshold accuracy characterizes robustness within different tolerance ranges. This multi-dimensional evaluation system can comprehensively reflect the comprehensive performance of ADDFNet in transparent object depth completion tasks.

4.4. Ablation Studies

To verify the effectiveness of each core design in the ADDFNet network proposed in this paper, we systematically studied the combination schemes and performance contributions of MDAM (containing DEDM and DSCA sub-components) and CMFR. By gradually removing or replacing these key components, we deeply analyzed the roles and synergistic effects of sub-components within MDAM and cross-modal fusion mechanisms in transparent object depth completion tasks, thereby providing empirical basis for the optimization design of network architecture.

 doi:10.20944/preprints202604.1051.v1

4.4.1. Overall Experiments

As shown in Figure 4, we quantitatively evaluate the contribution of each module through ablation studies.

1) **Baseline Model**: We removed the core innovative modules from the ADDFNet network, retaining only the basic encoder-decoder architecture, referred to as the baseline model.

2) **Baseline Model + Dilated Convolution Module**: To verify the effectiveness of MDAM, we introduced a dilated convolution module for comparison in the second experiment, with its parameter scale kept consistent with MDAM to ensure fair evaluation.

3) **Baseline Model + MDAM**: The third experiment adds MDAM to the baseline model configuration, containing DEDM, the DSCA unit, and feature projection mapping components.

4) **Baseline Model + MDAM + CMFR**: The fourth experimental variant further introduces the CMFR module on top of MDAM to enhance cross-modal feature information interaction capability.

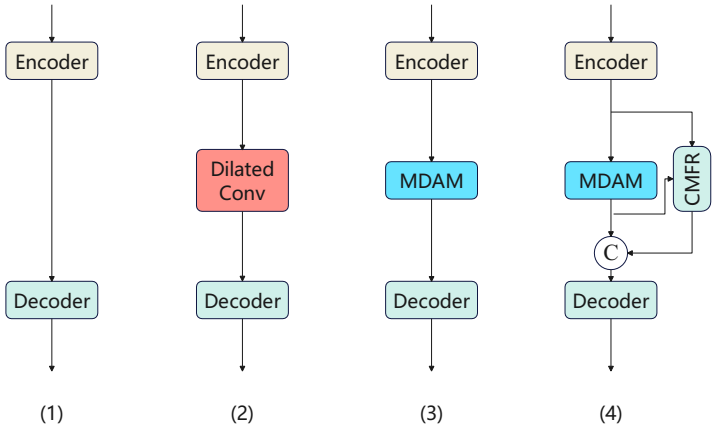

Figure 4. Overall arrangement of ablation study results.

To ensure fairness, all four modules replace the corresponding components in the backbone network.

To maintain consistency of experimental conditions and ensure fair comparison, all ablation studies are trained and tested on the TransCG dataset [17], using the same preprocessing pipeline, data augmentation strategies, and evaluation metrics as the main experiments. In terms of training settings, all models adopt the same optimization strategy as the main experiments: training with the Adam optimizer for 100K iterations, initial learning rate set to 1×10^{-3}, using a multi-step learning rate decay strategy, decaying the learning rate by a factor of 0.2 at 12.5%, 37.5%, 62.5%, and 87.5% of the total iteration process. Batch size is fixed at 16, and input image resolution is uniformly adjusted to 320x240 pixels. Under this unified training paradigm, performance comparisons between configurations can more objectively reflect the independent contributions and synergistic effects of each module, providing strong empirical basis for the rationality of network architecture design. It should be noted that for rapid testing, this section's experiments only conducted 100K iterations of training, so related results will be slightly lower than data obtained with more sufficient training settings in comparison experiment tables.

To provide an intuitive comparison of visual differences across configurations, Figure 5 presents qualitative ablation results, while quantitative metrics are summarized in Table 1. When only the basic encoder-decoder architecture is retained, the model obtains an RMSE of 0.055, indicating limited ability to capture fine contour details of transparent objects. After introducing a dilated convolution module into the baseline, the RMSE decreases to 0.043, showing moderate improvement but still inferior to the MDAM-based design. The Baseline + MDAM configuration further reduces RMSE to 0.040, demonstrating clear advantages. The MDAM module explicitly extracts spatial gradient cues through multi-directional differential convolution, and its internal DEDM and DSCA sub-components collaboratively perform local detail enhancement and global context modeling. Progressive comparisons

 doi:10.20944/preprints202604.1051.v1

of predicted depth maps clearly show that, compared with the baseline and the Baseline + Dilated Conv variant, introducing MDAM significantly improves the recovery of edge details and contour structures for transparent objects (e.g., syringes and plastic bottles), producing object shapes that better match real scenes and confirming the key role of MDAM in transparent-edge perception. Building on this, adding the cross-modal feature refinement (CMFR) module further aligns fine-grained depth details and regional depth consistency with ground truth (GT). The quantitative results also confirm this gain: the Baseline + MDAM + CMFR setting further reduces RMSE to 0.037. This indicates that the proposed adaptive fusion mechanism effectively suppresses feature degradation and strengthens the synergy among multiple information sources, including raw inputs, differential features, and attention features, thereby further improving depth estimation accuracy. Overall, the ablation study demonstrates strong complementarity among ADDFNet components: 1) as the core feature extraction unit, MDAM (with DEDM and DSCA) specifically enhances edge perception for transparent objects through multi-directional differential enhancement and dynamic-static attention collaboration, alleviating blurred boundaries and depth distortion; 2) CMFR optimizes the integration ratio among raw-input, differential, and attention features for efficient multimodal fusion, further improving robustness and precision. These qualitative and quantitative findings jointly validate the effectiveness of the proposed ADDFNet architecture.

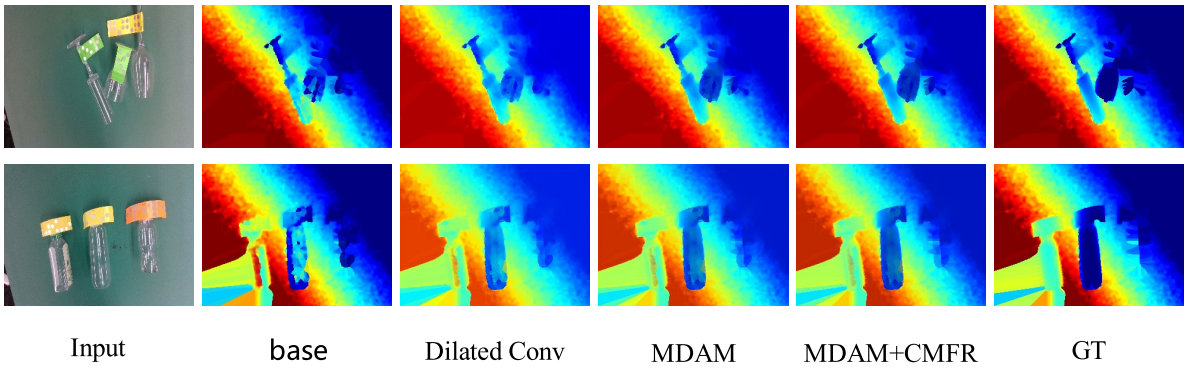

Figure 5. Qualitative analysis of ablation studies.

Table 1. Results of ablation studies on the TransCG dataset.

Configuration	Metrics					
	RMSE↓	REL↓	MAE↓	$\delta_{1.05}$ ↑	$\delta_{1.10}$ ↑	$\delta_{1.25}$ ↑
Baseline	0.055	0.080	0.040	55.42	71.14	85.37
Baseline + Dilated Conv	0.043	0.061	0.028	68.90	86.20	91.24
Baseline + MDAM	0.040	0.056	0.025	73.90	89.50	98.30
Baseline + MDAM + CMFR	0.037	0.052	0.022	77.20	91.80	98.60

4.4.2. Effectiveness of DEDM Sub-Component in MDAM

To verify the key role of DEDM, the differential enhancement sub-component within MDAM, in transparent object depth completion tasks, we conducted comparative analysis with various feature extraction modules with similar functional positioning. As shown in Table 2, while keeping other components of MDAM unchanged, we replaced DEDM with Fast Depth Completion for Transparent Objects [52], Transparent Object Depth Estimation with Transformer [53], Deformable Convolution for Depth Completion [54], and Visual Attention Network [55]. Experimental results show that DEDM, as a sub-component of MDAM, demonstrates significant advantages in all key metrics. To visually demonstrate the visual differences between configurations, Figure 6 provides qualitative analysis results of ablation studies.

 doi:10.20944/preprints202604.1051.v1

Table 2. Comparative ablation study of DEDM sub-component within MDAM and similar functional modules.

Module Configuration	Performance Metrics					
	RMSE↓	REL↓	MAE↓	$\delta_{1.05}$ ↑	$\delta_{1.10}$ ↑	$\delta_{1.25}$ ↑
DEDM	0.037	0.052	0.022	77.20	91.80	98.60
DEDM → FDCT [52]	0.046	0.065	0.030	68.50	85.20	97.50
DEDM → TODE-Trans [53]	0.045	0.064	0.029	69.60	86.00	97.70
DEDM → DCDC [54]	0.049	0.068	0.032	66.20	83.90	97.20
DEDM → VAN [55]	0.053	0.073	0.035	63.10	81.50	96.80

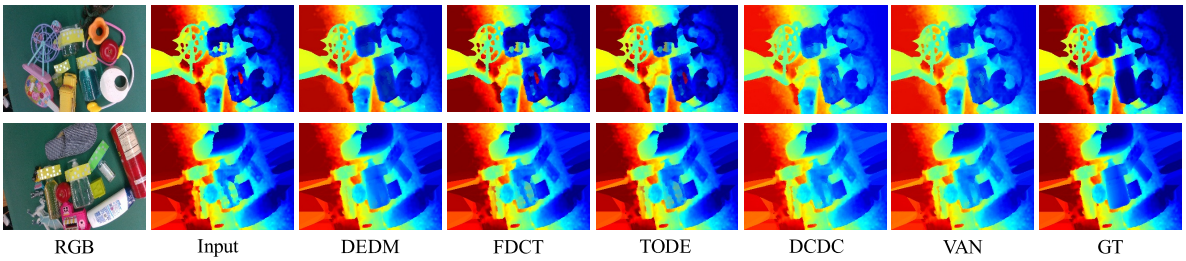

Figure 6. DEDM ablation study.

Specifically, when using Fast Depth Completion Network (FDCT) to replace DEDM, the model's RMSE is 0.046; Transparent Object Depth Estimation with Transformer (TODE-Trans) has an RMSE of 0.045. While both show some improvement, they still lag significantly behind DEDM. The Deformable Convolution for Depth Completion Module (DCDC) has an RMSE of 0.049, with its deformable convolution kernel having large computational overhead and limited stability. The Visual Attention Network (VAN) further increases RMSE to 0.053, with more obvious error accumulation in complex refraction and reflection scenes.

In contrast, the DEDM module achieves an RMSE of 0.037, demonstrating optimal performance. This advantage mainly stems from the dual design philosophy of DEDM: 1) Adaptively extracting spatial gradient information through learnable multi-directional differential convolution, achieving precise perception of transparent object edges; 2) Adopting a differential feature enhancement mechanism to strengthen the representation capability of gradient information at multiple scales. These results fully verify the effectiveness and superiority of DEDM in transparent object depth completion tasks, with its differential enhancement mechanism providing key technical support for handling visual perception problems with sparse edges susceptible to background interference.

4.4.3. Effectiveness of DSCA Sub-component in MDAM

To evaluate the DSCA sub-component within MDAM in transparent object depth completion tasks, we conducted comparative analysis with various attention mechanisms with similar functional positioning. As shown in Table 3, while keeping other components of MDAM unchanged, we replaced DSCA with Dynamic Convolution [37], Large Kernel Attention Module [41], Visual Attention Network [55], and Edge-Aware Spatial Propagation Network [56]. Experimental results show that DSCA, as a sub-component of MDAM, demonstrates significant advantages in all key metrics. To visually demonstrate the visual differences between configurations, Figure 7 provides qualitative analysis results of ablation studies.

 doi:10.20944/preprints202604.1051.v1

Table 3. Comparative ablation study of DSCA sub-component within MDAM and similar functional modules.

Module Configuration	Performance Metrics					
	RMSE↓	REL↓	MAE↓	$\delta_{1.05}$ ↑	$\delta_{1.10}$ ↑	$\delta_{1.25}$ ↑
DSCA	0.037	0.052	0.022	77.20	91.80	98.60
DSCA → DC [37]	0.054	0.074	0.035	62.80	80.90	96.50
DSCA → LKA [41]	0.051	0.071	0.033	64.60	82.50	96.90
DSCA → VAN [55]	0.046	0.065	0.030	68.90	85.80	97.60
DSCA → EASP [56]	0.048	0.067	0.031	67.00	84.30	97.30

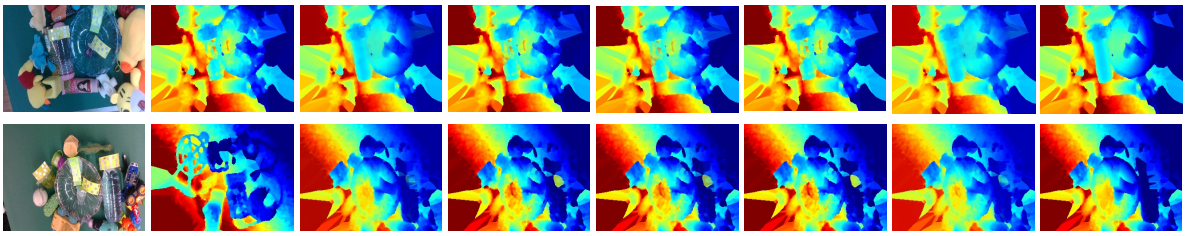

Figure 7. DSCA ablation study.

Specifically, when using Dynamic Convolution to replace DSCA, the model's RMSE is 0.054, showing the most obvious error increase, indicating insufficient robustness of early dynamic attention schemes in this task. The Large Kernel Attention Module has an RMSE of 0.051, showing improvement over early methods but still limited by static weights. The Visual Attention Network has an RMSE of 0.046, performing best among comparison methods but with higher computational complexity. The Edge-Aware Spatial Propagation Network has an RMSE of 0.048, still struggling to balance efficiency and accuracy.

Comprehensive comparative analysis shows that the DSCA module demonstrates excellent performance advantages in transparent object depth completion tasks, achieving an RMSE of 0.037, significantly outperforming other comparison methods. This performance advantage can be attributed to multiple innovations in the architectural design of the DSCA module:

First, DSCA adopts a dual-path design strategy of dynamic-static fusion, effectively balancing the needs of global modeling and local adaptation. The static large-kernel convolution branch provides stable long-range spatial interaction through fixed weights, ensuring coherent modeling of the overall geometric structure of transparent objects; while the dynamic depth convolution branch adaptively adjusts local response weights according to input features, achieving refined enhancement of object boundaries and internal details. This dual-path collaborative mechanism overcomes the limitations of single dynamic or static methods.

Second, DSCA achieves maximization of feature representation capability while ensuring computational efficiency through a lightweight architecture design of "channel splitting - parallel enhancement - feature recombination". The channel splitting strategy decomposes the feature flow into multiple subspaces, allowing different branches to focus on specific feature dimensions; the parallel enhancement mechanism enables static and dynamic paths to independently optimize their respective modeling capabilities; feature recombination achieves integration of complementary advantages through cross-path information fusion.

Experimental results show that the dynamic-static fusion mechanism of DSCA provides an effective solution for handling transparent object depth completion, a visual perception problem with spatial heterogeneity and background interference characteristics. Its design philosophy not only demonstrates superiority in this task but also provides valuable reference for network architecture design in other similar complex visual tasks.

 doi:10.20944/preprints202604.1051.v1

4.5. Comparison with SOTA Methods

In this section, based on the ClearPose and TransCG datasets, we conduct systematic comparative analysis between the proposed ADDFNet and other transparent object depth completion networks. Evaluation methods cover classic depth completion and transparent object-specific models, including Deep Depth Completion [19], NLSPN [23], Sparse-to-Dense [18], DistillGrasp [15], DualTransNet [29], and RGB-D Local Implicit Function [26]. Experimental evaluation uses ADDFNet as the baseline model. To ensure the credibility and consistency of comparison results, we adopt the following processing strategies for different benchmark methods: prioritize using officially released code and evaluation results for comparison; for algorithms lacking public implementations, retrain based on the same training data and experimental parameters to ensure fairness.

1) Quantitative Analysis: As shown in Table 4 and Table 5, ADDFNet demonstrates significant performance advantages on both ClearPose and TransCG datasets. On the TransCG dataset, ADDFNet's RMSE is 0.029, a 19.44% decrease compared to the second-best method RGB-D Local Implicit Function [26] (RMSE=0.036); on the ClearPose dataset, ADDFNet's RMSE is 0.110, a 20.29% decrease compared to the second-best method DualTransNet [29] (RMSE=0.138). Simultaneously, ADDFNet also achieves optimal results in the three threshold accuracy metrics $\delta_{1.05}$, $\delta_{1.10}$, and $\delta_{1.25}$, indicating that this method has better stability in transparent object edge and structure recovery.

Furthermore, the differences in improvement between the two datasets are closely related to their data distribution characteristics. Although relative reduction percentages are similar (19.44% vs. 20.29%), the absolute error reduction on ClearPose is larger (0.028 vs. 0.007). The main reason is that ClearPose adopts a "synthetic-real hybrid" construction with a smaller scale of real samples, resulting in more obvious domain appearance and noise statistical differences, causing baseline methods to be more prone to error accumulation in edge refraction regions; ADDFNet is more robust to such cross-domain disturbances through multi-directional differential enhancement and geometric consistency constraints, thus obtaining more significant absolute benefits. In contrast, the TransCG dataset has a larger scale and more comprehensive real scene coverage, with higher baseline performance for all methods, so improvements mainly manifest as stable optimization in low-error intervals.

Table 4. Performance on TransCG dataset.

Method	RMSE↓	REL↓	MAE↓	$\delta_{1.05}$ ↑	$\delta_{1.10}$ ↑	$\delta_{1.25}$ ↑
DeepDepthCompletion [19]	0.045	0.062	0.035	62.50	83.80	97.20
NLSPN [23]	0.041	0.058	0.031	66.25	85.90	96.60
Sparse-to-Dense [18]	0.047	0.064	0.037	61.80	84.50	97.10
DistillGrasp [15]	0.038	0.056	0.027	70.40	89.25	98.40
DualTransNet [29]	0.037	0.055	0.026	72.15	90.10	97.55
RGB-D Local Implicit [26]	0.036	0.054	0.024	73.25	90.65	98.70
Ours	0.029	0.045	0.016	78.83	94.74	99.15

Table 5. Performance on ClearPose dataset (Heavy occlusion scenario).

Method	RMSE↓	REL↓	MAE↓	$\delta_{1.05}$ ↑	$\delta_{1.10}$ ↑	$\delta_{1.25}$ ↑
DeepDepthCompletion [19]	0.175	0.090	0.068	35.40	70.25	94.85
NLSPN [23]	0.158	0.075	0.055	47.20	76.80	96.25
Sparse-to-Dense [18]	0.180	0.092	0.070	34.15	69.10	94.60
DistillGrasp [15]	0.142	0.062	0.046	55.80	82.40	97.15
DualTransNet [29]	0.138	0.060	0.044	58.90	83.95	97.50
RGB-D Local Implicit [26]	0.146	0.064	0.048	53.75	81.20	96.95
Ours	0.110	0.033	0.028	76.33	93.84	98.88

2) Qualitative Analysis: Qualitative visualizations compare DeepDepthCompletion [19], NLSPN [23], Sparse-to-Dense [18], DistillGrasp [15], DualTransNet [29], RGB-D Local Implicit [26], and our ADDFNet on the ClearPose and TransCG datasets. Visual comparisons on the TransCG dataset are

 doi:10.20944/preprints202604.1051.v1

shown in Fig. 8, and those on the ClearPose dataset are shown in Fig. 9. Odd and even rows respectively present qualitative depth estimation comparisons of transparent objects under different scenes. From left to right, each sample shows the RGB input, initial depth input, predictions from representative baselines , the reconstruction result of our method , and the ground-truth depth. By comparing baseline predictions and GT, it can be directly observed whether our method accurately recovers the geometric structure and depth variations of transparent objects in standard scenes. Experimental results show that ADDFNet exhibits clearer boundary recovery and more complete internal detail reconstruction on both datasets. Consistent with quantitative results in Tables 4 and 5, ADDFNet has obvious advantages in edge clarity, internal structure completeness, and noise suppression.

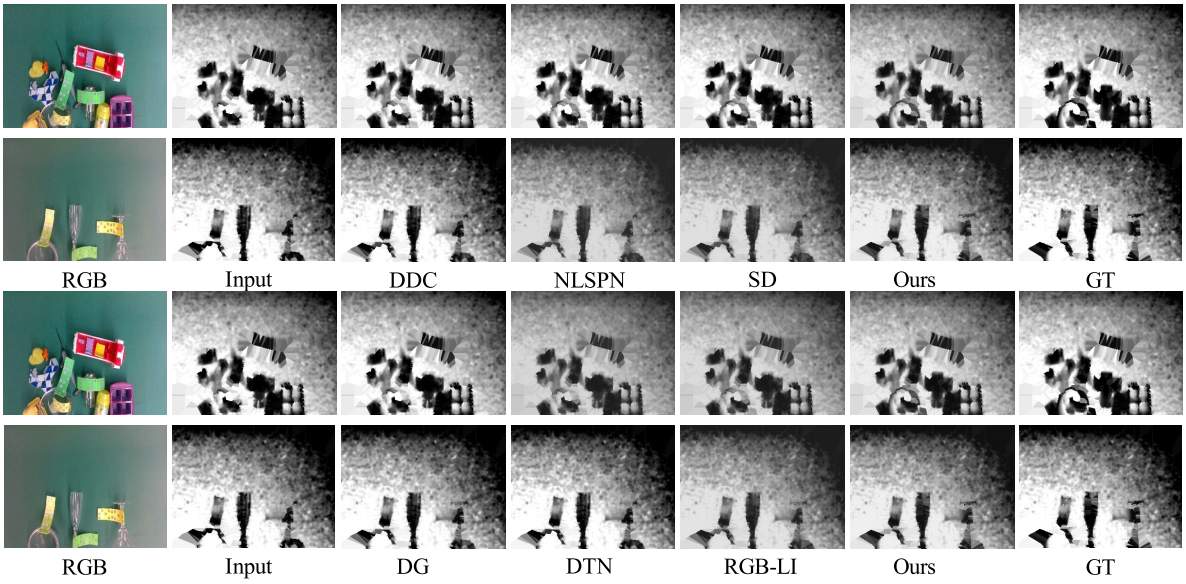

Figure 8. Comparative experiments on the TransCG dataset.

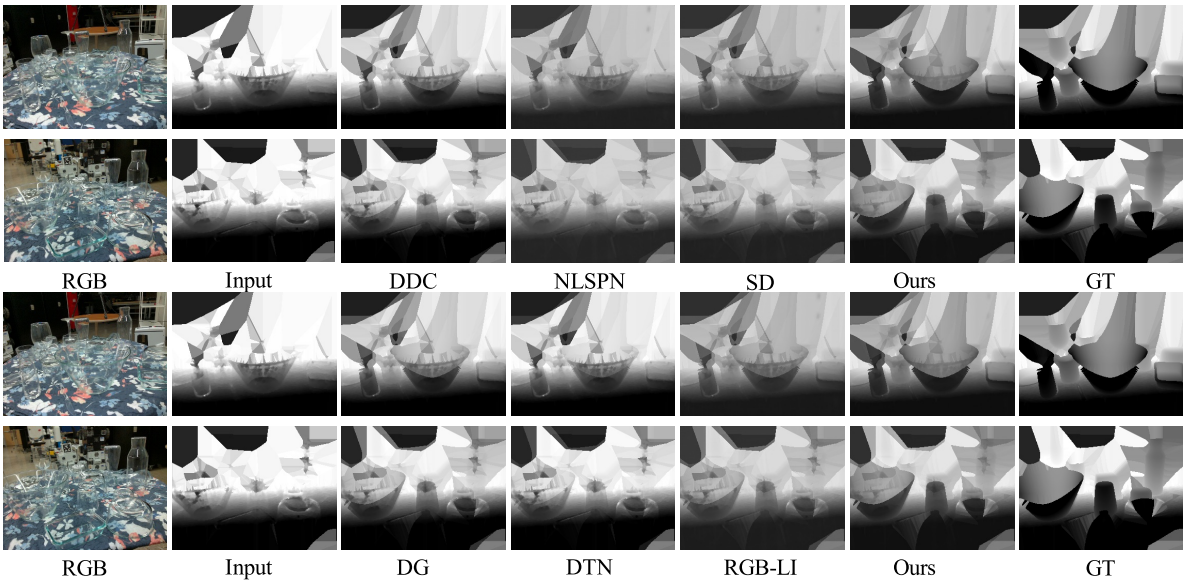

Figure 9. Comparative experiments on the ClearPose dataset (Heavy occlusion scenario).

The performance advantages of ADDFNet can be attributed to multiple innovations in its architectural design, which have been fully verified in the experimental results of Tables 4 and 5: First, MDAM effectively strengthens the network's perception capability for fine contours of transparent objects by explicitly modeling spatial gradient variations. This is directly reflected in ADDFNet's excellent performance in the edge-sensitive metric $\delta_{1.05}$, showing significant improvement compared to traditional methods like DeepDepthCompletion . Compared to RGB-D Local Implicit, ADDFNet performs

 doi:10.20944/preprints202604.1051.v1

better in edge detail recovery, mainly benefiting from the multi-directional differential enhancement mechanism of the DEDM sub-component within MDAM, which can more accurately capture gradient variations at transparent object edges. Second, the DSCA sub-component within MDAM adopts a dual-path design strategy of dynamic-static fusion, balancing the needs of global modeling and local adaptation, achieving adaptive modeling of complex depth distributions in transparent regions. This explains ADDFNet's leading advantages in RMSE and MAE metrics. Finally, CMFR improves the robustness of multi-source feature collaborative modeling by recalibrating and enhancing depth features layer by layer using RGB context during the encoding stage, ensuring stable performance of ADDFNet across different datasets. The synergistic effect of these innovative mechanisms enables ADDFNet to achieve stable and consistent performance improvement in transparent object depth completion tasks, outperforming comparison methods on both challenging TransCG and ClearPose datasets.

5. Conclusion

This paper addresses the problem of missing depth information for transparent objects. Our proposed ADDFNet network enhances depth completion performance through two key designs: MDAM and CMFR. As the core feature enhancement module, MDAM achieves local gradient detail reinforcement and global-local adaptive modeling through its internal DEDM and DSCA sub-components, effectively improving transparent object contour and structure recovery capabilities. CMFR enhances the efficiency and robustness of multi-source feature fusion by leveraging RGB context to perform layer-wise enhancement of depth feature representations during the encoding stage. Evaluation results on the ClearPose and TransCG datasets demonstrate that ADDFNet outperforms comparison methods on metrics including RMSE, REL, MAE, and threshold accuracy, while exhibiting better stability in edge recovery and detail reconstruction. It should be noted that the current method still has two limitations: first, modules such as MDAM introduce additional computational overhead, leaving room for optimization in deployment efficiency on resource-constrained platforms; second, the model's generalization capability still requires further improvement when facing high-curvature, strongly reflective, or extremely thin transparent structures. Based on these limitations, subsequent work will focus on lightweight design and robustness enhancement for complex scenarios.

6. Future Work

Addressing the aforementioned limitations, future research will focus on developing lightweight attention mechanisms to reduce model complexity, exploring geometry-aware data augmentation techniques to enhance generalization capabilities for diverse transparent shapes and materials, and validating the method through additional datasets and real robotic platforms to enhance its practicality. Furthermore, we plan to extend ADDFNet to broader robotic vision tasks, such as transparent object pose estimation and scene understanding. This will not only help validate the universality and practical value of our method but also provide broader technical support for related research fields.

Author Contributions: Conceptualization, N.L. and J.W.; Methodology, N.L. and Y.L.; Software, N.L.; Validation, N.L., Y.L. and Y.W.; Formal analysis, N.L. and X.Y.; Investigation, N.L. and Y.L.; Resources, N.L. and Y.W.; data curation, N.L. and X.Y.; Writing—original draft preparation, N.L.; Writing—review and editing, N.L., Y.L. and X.Y.; Visualization, N.L.; Supervision, N.L. and Y.L.; Project administration, N.L. and Y.L.; Funding acquisition, N.L., Y.L. and X.Y. All authors have read and agreed to the published version of the manuscript.

Funding: The study was funded by Zhangzhou Municipal Natural Science Foundation (ZZ2025JH08)

Data Availability Statement: The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author.

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

 doi:10.20944/preprints202604.1051.v1

References

1. Bicchi, A.; Kumar, V. Robotic grasping and contact: A review. In Proceedings of the Proceedings 2000 ICRA. Millennium conference. IEEE international conference on robotics and automation. Symposia proceedings (Cat. No. 00CH37065). IEEE, 2000, Vol. 1, pp. 348–353.
2. Redmon, J.; Angelova, A. Real-time grasp detection using convolutional neural networks. *IEEE International Conference on Robotics and Automation (ICRA)* **2015**, pp. 1316–1322.
3. Qian, Y.; Gong, M.; Yang, Y.H. Transparent object reconstruction based on spatio-temporal light transport analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2020**, *42*, 3060–3073.
4. Xie, E.; Wang, W.; Wang, W.; Ding, M.; Shen, C.; Luo, P. Segmenting Transparent Objects in the Wild. In Proceedings of the Proceedings of the European Conference on Computer Vision (ECCV), 2020, pp. 696–711.
5. Schwarz, M.; Milan, A.; Selvam Periyasamy, A.; Behnke, S. RGB-D object detection and semantic segmentation for autonomous manipulation in clutter. *The International Journal of Robotics Research* **2017**, *37*, 437–451. https://doi.org/10.1177/0278364917713117.
6. Qi, C.R.; Yi, L.; Su, H.; Guibas, L.J. PointNet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in Neural Information Processing Systems* **2017**, *30*, 5099–5108.
7. Sajjan, S.; Moore, M.; Pan, M.; Nagaraja, G.; Lee, J.C.; Zeng, A.; Song, S. ClearGrasp: 3D Shape Estimation of Transparent Objects for Manipulation. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), 2020, pp. 3634–3642. https://doi.org/10.1109/ICRA40945.2020.9197518.
8. Sajjan, S.; Moore, M.; Pan, M.; Nagaraja, G.; Lee, J.; Zeng, A.; Song, S. Learning to see transparent objects. *IEEE International Conference on Robotics and Automation (ICRA)* **2020**, pp. 1–8.
9. He, K.; Sun, J.; Tang, X. Single image haze removal using dark channel prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2011**, *33*, 2341–2353.
10. Zhu, Q.; Mai, J.; Shao, L. A fast single image haze removal algorithm using color attenuation prior. *IEEE Transactions on Image Processing* **2015**, *24*, 3522–3533.
11. Nayar, S.K.; Ikeuchi, K.; Kanade, T. Shape from interreflections. *International Journal of Computer Vision* **1991**, *6*, 173–195.
12. Kutulakos, K.N.; Steger, E. A theory of refractive and specular 3D shape by light-path triangulation. *International Journal of Computer Vision* **2008**, *76*, 13–29.
13. Chen, X.; Zhang, H.; Yu, Z.; Opipari, A.; Jenkins, O.C. ClearPose: Large-scale Transparent Object Dataset and Benchmark. In Proceedings of the European Conference on Computer Vision, 2022.
14. 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 Proceedings of Machine Learning Research, 2021, Vol. 164, pp. 827–838.
15. 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* **2024**, *9*, 8945–8952.
16. Hua, Z.; Yu, S.; Wang, W.; Guan, Y.; Xia, Y. TCRNet: Transparent Object Depth Completion With Cascade Refinements. *IEEE Transactions on Automation Science and Engineering* **2024**.
17. Fang, H.S.; Wang, J.; Gou, Y.; Lu, H. TransCG: A Large-Scale Real-World Dataset for Transparent Object Depth Completion and A Grasping Baseline. *IEEE Robotics and Automation Letters* **2022**, *7*, 7383–7390.
18. Ma, F.; Karaman, S. Sparse-to-Dense: Depth Prediction from Sparse Depth Samples and a Single Image. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2018, pp. 479–487.
19. Zhang, Y.; Funkhouser, T. Deep Depth Completion of a Single RGB-D Image. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 175–185.
20. Cheng, X.; Wang, P.; Yang, R. Depth Estimation via Affinity Learned with Convolutional Spatial Propagation Network. In Proceedings of the Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 103–119.
21. Tang, J.; Tian, F.P.; Feng, W.; Li, J.; Tan, P. GuideNet: Guided Anisotropic Diffusion for Depth Completion. In Proceedings of the Proceedings of the European Conference on Computer Vision (ECCV), 2020.
22. Hu, M.; Wang, S.; Li, B.; Ning, S.; Fan, L.; Gong, X. FuseNet: An RGB-D Fusion Architecture for Depth Completion. *IEEE Robotics and Automation Letters* **2019**, *4*, 4424–4431.
23. Park, J.; Joo, K.; Hu, Z.; Liu, C.K.; So Kweon, I. Non-local Spatial Propagation Network for Depth Completion. *European Conference on Computer Vision (ECCV)* **2020**, pp. 120–136.

 doi:10.20944/preprints202604.1051.v1

24. Li, X.; Liu, Y.; Chen, X.; Wang, Y. ACMNet: Adaptive Context-Aware Multi-Scale Network for Depth Completion. *IEEE Transactions on Image Processing* **2021**.
25. Hu, M.; Wang, S.; Li, B.; Ning, S.; Fan, L.; Gong, X. PENet: Towards Precise and Efficient Image Guided Depth Completion. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), 2021.
26. Zhu, L.; Mousavian, A.; Xiang, Y.; Mazhar, H.; van Eenbergen, J.; Debnath, S.; Fox, D. RGB-D Local Implicit Function for Depth Completion of Transparent Objects. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 4647–4656.
27. Ku, J.; Harakeh, A.; Waslander, S.L. DeepLiDAR: Deep Surface Normal Guided Depth Prediction for Outdoor Scene From Sparse LiDAR Data and Single Color Image. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3313–3322.
28. Yan, Z.; Zhang, R.; Wang, X. PENet: Penetration-Aware Depth Completion for Transparent Objects. *arXiv preprint arXiv:1907.00000* **2019**.
29. Liu, B.; Li, H.; Wang, Z.; Xue, T. Transparent Depth Completion Using Segmentation Features. *ACM Transactions on Multimedia Computing, Communications, and Applications* **2024**, *20*, 373:1–373:19. DualTransNet is proposed in this work.
30. Chen, Z.; Liu, Y.; Wang, P. SRNet-Trans: A Signal-Image Guided Depth Completion Regression Network for Transparent Object. *Applied Sciences* **2023**, *15*, 10566.
31. Li, J.; Zhang, R.; Wang, X. Consistent Depth Prediction for Transparent Object Reconstruction from RGB-D Camera. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2023, pp. 4567–4576.
32. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* **2018**, pp. 7132–7141.
33. Woo, S.; Park, J.; Lee, J.Y.; So Kweon, I. Convolutional block attention module. *Proceedings of the European Conference on Computer Vision (ECCV)* **2018**, pp. 3–19.
34. Su, Z.; Liu, W.; Yu, Z.; Hu, D.; Liao, Q.; Tian, L.; Pietikainen, M.; Liu, L. Pixel difference networks for efficient edge detection. *International Conference on Computer Vision (ICCV)* **2021**, pp. 5117–5127.
35. Su, Z.; Liu, W.; Yu, Z.; Hu, D.; Liao, Q.; Tian, L.; Gao, X.; Pietikainen, M. Pixel difference convolutional neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2023**, *45*, 4296–4310.
36. Duan, K.; Bai, S.; Xie, L.; Qi, H.; Huang, Q.; Tian, Q. CenterNet: Keypoint Triplets for Object Detection. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 6569–6578.
37. Chen, Y.; Dai, X.; Liu, M.; Chen, D.; Yuan, L.; Liu, Z. Dynamic convolution: Attention over convolution kernels. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* **2020**, pp. 11030–11039.
38. Sobel, I.; Feldman, G. A 3x3 isotropic gradient operator for image processing. *A Talk at the Stanford Artificial Intelligence Project* **1968**.
39. Ding, X.; Zhang, X.; Han, J.; Ding, G. Scaling up your kernels to 31x31: Revisiting large kernel design in CNNs. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* **2022**, pp. 11963–11975.
40. Ranftl, R.; Bochkovskiy, A.; Koltun, V. Vision transformers for dense prediction. *Proceedings of the IEEE/CVF International Conference on Computer Vision* **2021**, pp. 12179–12188.
41. Ding, X.; Zhang, X.; Han, J.; Ding, G. Scaling up your kernels to 31x31: Revisiting large kernel design in CNNs. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* **2022**, pp. 11963–11975.
42. Liu, Z.; Mao, H.; Wu, C.Y.; Feichtenhofer, C.; Darrell, T.; Xie, S. ConvNeXt: A pure ConvNet for the 2020s. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* **2022**, pp. 11976–11986.
43. Yang, B.; Bender, G.; Le, Q.V.; Ngiam, J. CondConv: Conditionally parameterized convolutions for efficient inference. *Advances in Neural Information Processing Systems* **2019**, *32*.
44. Chen, Z.; He, Y.; Shiri, I.; Li, H.; Li, Y.; Zhang, J.; Guo, X.; Li, H.; Wang, Q. DEA-Net: Single image dehazing based on detail-enhanced convolution and content-guided attention. *IEEE Transactions on Image Processing* **2024**, *33*, 1002–1015.
45. Ding, X.; Zhang, X.; Ma, N.; Han, J.; Ding, G.; Sun, J. RepVGG: Making VGG-style ConvNets Great Again. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* **2021**, pp. 13733–13742.

 doi:10.20944/preprints202604.1051.v1

46. Cohen, T.; Welling, M. Group equivariant convolutional networks. *International Conference on Machine Learning* **2016**, pp. 2990–2999.
47. Johnson, J.; Alahi, A.; Fei-Fei, L. Perceptual losses for real-time style transfer and super-resolution. *Proceedings of the European Conference on Computer Vision (ECCV)* **2016**, pp. 694–711.
48. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *International Conference on Learning Representations (ICLR)* **2015**.
49. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. ImageNet: A large-scale hierarchical image database. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* **2009**, pp. 248–255.
50. Eigen, D.; Puhrsch, C.; Fergus, R. Depth map prediction from a single image using a multi-scale deep network. *Advances in Neural Information Processing Systems* **2014**, *27*, 2366–2374.
51. Gao, H.; Liu, X.; Qu, M.; Huang, S. PDANet: Self-supervised monocular depth estimation using perceptual and data augmentation consistency. *Applied Sciences* **2021**, *11*, 5383.
52. Li, T.; Chen, Z.; Liu, H.; Wang, C. FDCT: Fast Depth Completion for Transparent Objects. *IEEE Robotics and Automation Letters* **2023**, *8*, 5823–5830.
53. Wang, S.; Zhu, X.; Zhang, Y.; Li, S.; Liu, Y.; Wang, J. TODE-Trans: Transparent Object Depth Estimation with Transformer. *IEEE International Conference on Robotics and Automation (ICRA)* **2023**, pp. 1–8.
54. Sun, X.; Ponce, J.; Wang, Y.X. Revisiting deformable convolution for depth completion. *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* **2023**, pp. 1234–1245.
55. Guo, M.H.; Lu, C.Z.; Liu, Z.N.; Hu, M.M.; Zhang, T.J.; Hu, S.M. Visual Attention Network. *Computational Visual Media* **2023**, *9*, 733–752.
56. Zhang, S.; Xu, S.; Li, X.; Yang, J. Edge-aware spatial propagation network for multi-view depth estimation. *Neural Processing Letters* **2023**, *55*, 1–20.