WEP Warwick Evans Publishing

International Journal of Computer Science and Information Technology

ISSN: 3005-9682 (Print), ISSN: 3005-7140 (Online) | Volume 8, Number 1, Year 2026
DOI: https://doi.org/10.62051/ijcsit.v8n1.03
Journal homepage: https://wepub.org/index.php/IJCSIT/index

# Research on a Curvature-Enhanced and Synergistic Attention-Based Multi-Task Perception Method for Transparent Objects

Jiajin Han [1, 2, 3], Sanpeng Deng [1, 2, 3, *], Yuming Qi [1, 2, 3], Xiumin Shi [1, 2, 3]

[1] Tianjin University of Technology and Education, Tianjin 300222, China

[2] Tianjin Key Laboratory of Intelligent Robot Technology and Application, Tianjin 300350, China

[3] Tianjin Bonus Robotics Technology Co., Ltd, Tianjin 300350, China

*Corresponding Author: Sanpeng Deng

## ABSTRACT

Transparent objects challenge monocular perception due to refraction, reflection, and weak textures, which hinder accurate depth estimation and segmentation. To overcome these issues, we propose CESINet, a curvature-enhanced synergistic attention network for transparent object perception. CESINet explicitly incorporates surface curvature as a high-order geometric prior to strengthen spatial representation and introduces a curvature-guided synergistic attention module to enable effective cross-task feature interaction between depth and segmentation branches. A curvature consistency loss further enforces geometric coherence across predictions. Experiments on the ClearPose dataset show that CESINet achieves 94.33% mIoU and 98.27% mAP for segmentation, improving over the multi-task baseline ISGNet by 1.49% and 0.44%, respectively. For depth estimation, CESINet attains an RMSE of 0.112 and REL of 0.060, reducing errors by 8.9% and 11.8% compared with the baseline. Ablation results demonstrate that removing curvature priors or attention modules leads to performance drops of up to 3.5% in segmentation and 12% in depth accuracy, confirming the complementary benefits of explicit geometry and synergistic learning. Overall, CESINet enhances geometric consistency and boundary sharpness while maintaining computational efficiency, providing a unified and scalable framework for multi-task transparent object understanding.

## KEYWORDS

Transparent object perception; Multi-task learning; Curvature prior; Depth estimation; Semantic segmentation; Attention mechanism

## 1. INTRODUCTION

Transparent objects pose persistent challenges for detection, segmentation, and geometric estimation due to appearance–background coupling induced by refraction and weak texture, leading to unstable boundaries and ambiguous shape reconstruction. Chen et al. introduced TOM-Net, which formulates transparent object matting as a refractive-flow estimation problem. Their two-stage network jointly regresses object masks, attenuation, and refractive fields from a single image, thus incorporating the coupling between transmission and geometry into an end-to-end learning framework [1]. Subsequently, Sajjan et al. proposed ClearGrasp, which performs synthetic-to-real mixed training to jointly infer surface normals and depth, effectively correcting missing and distorted depth for transparent objects and embedding explicit geometric constraints into downstream grasping and 3D reconstruction [2].

To enable more challenging segmentation benchmarks, Xie et al. released the Trans10K dataset and developed the boundary-aware TransLab, emphasizing the significance of boundary cues for transparent region segmentation [3]. In parallel, Kalra et al. incorporated polarization cues into deep networks, demonstrating that multimodal physical imaging effectively decouples transparent regions from their backgrounds [4]. Fang et al. further published the TransCG dataset [5], while Wang et al. proposed MVTrans, exploiting multi-view information to enhance geometric consistency and registration quality [6]. In broader 3D vision contexts, Hamdi et al. introduced MVTN, a multi-view transformation network that learns viewpoint transformations to substantially improve 3D understanding, validating the efficacy of multi-view features in complex scene modeling [7].

At the feature modeling level, attention mechanisms provide an effective dynamic reweighting strategy for scenes characterized by weak textures, strong boundaries, and sparse geometry. Woo et al. proposed the Convolutional Block Attention Module (CBAM), which has demonstrated the effectiveness of channel–spatial attention across various vision tasks [8]. In the field of visual question answering, Lu et al. introduced co-attention mechanisms, enabling bidirectional guidance between modalities for semantic alignment [9], an idea later extended to visual multi-task learning. Among these extensions, Yu et al. proposed Multidimensional Collaborative Attention (MCA) [10], while Cui et al. incorporated collaborative multi-task structures to mitigate negative transfer in unified frameworks [11]. Similarly, Misra et al. developed the Cross-Stitch Unit, which adaptively balances shared and task-specific representations through linear combinations across tasks [12], providing a foundation for synergistic attention mechanisms in transparent object perception. Overall, despite progress from both the data and modeling perspectives, the exploitation of high-order geometric priors remains insufficient.

To address the aforementioned limitations, Liu et al. [18] proposed a unified framework for monocular depth estimation and transparent object segmentation, incorporating an iterative semantic–geometric fusion mechanism that achieved promising results on both tasks. This work established a deep coupling between semantic and geometric representations in transparent scenes, significantly improving perception accuracy. However, it did not explicitly exploit high-order geometric priors such as surface curvature. As an intrinsic descriptor of local shape, curvature exhibits strong stability independent of texture and illumination variations. Recent studies have shown that integrating curvature as an explicit geometric constraint can substantially enhance surface modeling and perception performance. For instance, SR-CurvANN demonstrated the advantages of curvature-driven surface reconstruction [13]; Harrison et al. revealed that using curvature as an input feature markedly improves segmentation and classification performance [14]; and the CFPS method leveraged curvature-guided point cloud sampling to achieve higher accuracy in classification and segmentation tasks [15]. More recent research has further confirmed that Gaussian curvature serves as a strong prior for stereo matching and depth estimation [16]. Nevertheless, most existing methods for transparent object perception rely primarily on gradient or normal constraints [17], without explicitly incorporating curvature information to enhance local geometric modeling.

Building upon the framework proposed by Liu et al. [18], this paper further introduces a curvature-enhanced geometric prior and proposes an improved method for transparent object perception. The main contributions of this work are summarized as follows:

(1) We propose a novel approach that explicitly incorporates surface curvature as a high-order geometric prior directly into a deep neural network, enabling the model to exploit curvature-aware geometric cues during learning.

(2) We design a curvature-enhanced synergistic attention mechanism that facilitates efficient geometric information exchange between multiple tasks, thereby improving cross-task feature interaction and semantic consistency.

(3) We introduce a curvature consistency loss ($L_{curv}$) into a hybrid loss formulation, ensuring that predictions not only approximate ground truth at the pixel level but also maintain structural coherence

in three-dimensional geometry. This design produces sharper segmentation boundaries and more accurate depth estimation results.

## 2. PROPOSED METHOD

This paper proposes a Curvature-Enhanced Synergistic Iterative Network (CESINet), an end-to-end multi-task learning framework designed to simultaneously address transparent object segmentation and depth estimation. Unlike existing methods, CESINet explicitly integrates curvature-derived geometric priors and employs an iterative decoding mechanism to achieve deep cross-task interaction. By introducing geometric constraints between the segmentation and depth branches, CESINet effectively alleviates the challenges of boundary ambiguity and depth inconsistency in transparent regions, thereby achieving synergistic optimization of both tasks.

### 2.1. Overall Framework

As illustrated in Figure 1, CESINet consists of four main components: an encoder, parallel feature streams for segmentation and depth, a curvature prior branch, and an iterative decoder. The input RGB image is first processed by a Vision Transformer (ViT) encoder to extract multi-scale contextual features. Then, two task-specific feature streams are constructed to learn representations for segmentation and depth estimation, respectively. Meanwhile, curvature features computed from the depth map serve as explicit geometric priors, which are injected into the decoding process to enhance 3D structural reasoning. The iterative decoder progressively fuses segmentation, depth, and curvature features across multiple scales, refining predictions through several iterations to produce both a semantic segmentation mask and a continuous depth map. This unified design enables CESINet to perform geometry-constrained multi-task modeling within a single framework.

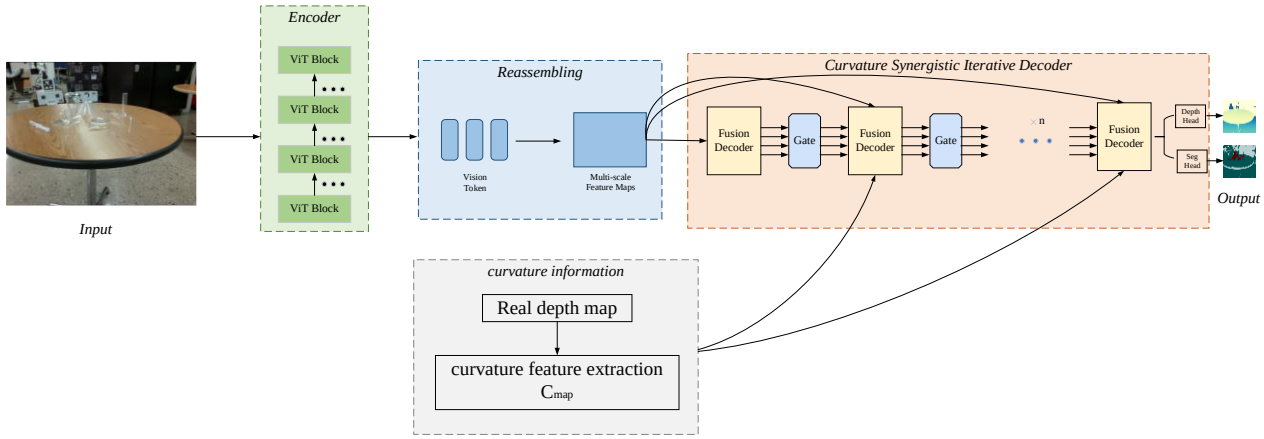


**Figure 1.** Overall architecture of the proposed CESINet framework

#### 2.1.1. Backbone Network Module

CESINet adopts the Vision Transformer (ViT) as its backbone feature extraction network. Originally proposed by Dosovitskiy et al. [19], ViT divides the input image into fixed-size patches, flattens each patch into a vector, and processes the resulting sequence through a standard Transformer encoder for global representation learning. Compared with conventional Convolutional Neural Networks (CNNs), which are limited by local receptive fields, the Transformer architecture leverages self-attention mechanisms to capture long-range dependencies across image regions. This enables ViT to better model complex structures and contextual relationships, which is particularly beneficial in transparent object perception, where scenes often exhibit blurred textures, weak edges, and strong contextual correlations. Consequently, ViT serves as a robust backbone that provides rich multi-scale contextual features for both segmentation and depth estimation tasks in CESINet.

### 2.1.2. Backbone Network Module

Since the Vision Transformer encoder outputs a one-dimensional sequence of tokens without explicit spatial structure, it is not directly suitable for dense prediction tasks. To address this issue, CESINet introduces a Reassemble Module, which reconstructs the spatial layout of token embeddings and transforms them into two-dimensional feature maps. Following the methodology proposed in DPT [20], this module establishes task-specific multi-scale feature pyramids for both depth estimation and semantic segmentation.

Specifically, the reassembly process involves two main steps:

(1) Feature Reshaping:

For the token sequences output from each selected layer of the encoder, the module reshapes them into two-dimensional feature maps with spatial dimensions (h, w), thereby restoring the spatial correspondence lost during the tokenization process.

(2) Parallel Pyramid Construction:

The module then constructs two parallel multi-scale feature pyramids for the subsequent tasks—depth estimation and semantic segmentation. The reshaped feature maps are passed through a projection layer (typically implemented as a 1×1 convolution) to generate two independent feature branches:

$$P_d = F_{d1}, F_{d2}, F_{d3}, F_{d4}$$

$$P_s = F_{s1}, F_{s2}, F_{s3}, F_{s4}$$

In both pyramids, the feature map resolution decreases progressively while the channel dimension increases, forming a coarse-to-fine hierarchy. This design provides rich multi-scale contextual information and ensures that both tasks receive well-aligned and semantically consistent feature representations for the subsequent iterative decoding process.

### 2.1.2. Curvature Synergistic Iterative Decoder

The Curvature Synergistic Iterative Decoder (CSID) is designed to progressively refine the predictions generated by the encoder through an iterative optimization and collaborative fusion strategy. In complex transparent object scenes, single-pass decoding frameworks often struggle to achieve structurally consistent results. For example, although TransDepth [21] introduces global contextual modeling via Transformers, its one-time fusion strategy still leads to depth misalignment and blurred boundaries around reflective or refractive regions. To overcome this limitation, CESINet employs a multi-scale, coarse-to-fine decoding strategy with N iterative refinements.

Each iteration in CSID consists of four decoding stages, gradually upsampling features from the lowest to the highest resolution. The key mechanism enabling effective fusion is the Curvature-Integrated Synergistic Channel–Spatial Attention (CI-SCSA) Fusion, which operates at every decoding scale. The iterative design ensures that each stage benefits from both the historical information of previous iterations and the current geometric features, thereby refining fine-grained structures such as object edges and transparent boundaries.

The iterative optimization process includes two major components:

Iterative Refinement Strategy:

The decoder does not complete prediction in a single forward pass but repeats the process N times. During the n-th iteration, it receives not only the reassembled features from the encoder but also the refined features from the (n−1)-th iteration. A lightweight Gated Unit regulates the integration between historical and current information, enforcing a gradual transition from coarse to fine

representations. This iterative refinement enables the network to focus on detailed regions progressively, leading to sharper segmentation and smoother depth surfaces.

Synergistic Fusion Mechanism:

At each scale of every iteration, a Synergistic Channel–Spatial Attention (SCSA) module replaces the conventional feature concatenation operation [22]. The SCSA module performs dynamic reweighting across spatial and channel dimensions to facilitate cross-task information exchange. It takes three inputs: (1) the depth feature map $F_d$; (2) the segmentation feature map $F_s$; and (3) the projected curvature map $C_{map}$.Through spatial–channel attention computation, SCSA adaptively assesses the importance of each feature source and fuses them accordingly. The segmentation branch provides semantic region cues that guide smooth surface reconstruction in the depth branch, while the curvature map contributes explicit geometric priors that enhance boundary sharpness and geometric consistency. Consequently, CSID achieves deep synergy among semantic, geometric, and curvature-aware representations, leading to more accurate and structurally coherent predictions.

# 3. CURVATURE-GUIDED GEOMETRIC MODELING

## 3.1. Curvature Feature Extraction from Depth Maps

Transparent objects often lack distinctive texture cues; thus, accurate perception in terms of depth estimation and semantic segmentation largely depends on geometric information, such as object shape and surface curvature. Traditional approaches typically rely on dense point cloud data to compute curvature; however, in monocular depth estimation scenarios, curvature features must be extracted directly from the predicted or intermediate depth maps.

During depth map acquisition, the raw sensor outputs frequently contain noise, missing values, and out-of-range measurements, which can significantly degrade geometric reliability. Therefore, a series of preprocessing steps are applied to enhance data quality before curvature computation. First, invalid or missing depth values are detected and corrected to avoid numerical bias in subsequent calculations. Then, median filtering and bilateral filtering are performed to remove local noise while preserving critical edge structures. For depth values exceeding the sensor's valid measurement range, clipping or masking operations are applied to constrain them within a physically reasonable domain. Subsequently, the preprocessed depth data are normalized to a predefined range to ensure compatibility with standard image processing pipelines. Finally, the normalized depth values are mapped to grayscale or pseudo-color images, providing an intuitive visualization of depth information. The overall preprocessing workflow is illustrated in Figure 2.

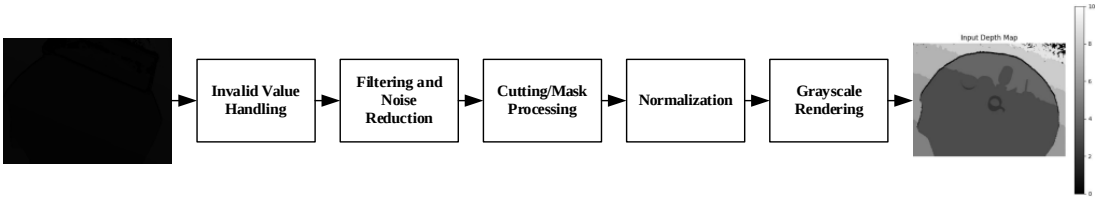


**Figure 2.** Preprocessing pipeline for raw depth map denoising and normalization

We propose a method to derive curvature features directly from two-dimensional depth maps, where the depth map can be regarded as a height field z=D(x, y). The curvature of a surface at a given point describes the degree of local bending of that surface. The two primary curvature measures—mean curvature and Gaussian curvature—can be computed from the first- and second-order partial derivatives of the depth function D with respect to the image coordinates x and y.

By treating the depth map as a Monge patch, represented as (x, y, D (x, y)), the mean curvature $H_c$ and Gaussian curvature $K_c$ of the surface can be calculated using the following formulations:

$$K_c = \frac{Z_{xx}Z_{yy} - Z_{xy}^2}{(1 + Z_x^2 + Z_y^2)^2}$$

$$H_c = \frac{(1 + Z_x^2)^2 Z_{yy} - 2Z_x Z_y Z_{xy} + (1 + Z_y^2)^2 Z_{xx}}{2(1 + Z_x^2 + Z_y^2)^{3/2}}$$

Considering that the two curvature measures described above are sensitive to noise, an additional and more robust curvature descriptor—the Laplacian operator of the depth map—is introduced:

$$\nabla^2 Z = Z_{xx} + Z_{yy}$$

The Laplacian operator, $\nabla^2 Z$, provides a scalar value at each pixel, indicating the local convexity or concavity of the surface at that point. By concatenating the three curvature representations described above—mean curvature, Gaussian curvature, and Laplacian curvature—a multi-channel curvature feature map, $C_{map} \in R^{3\times H\times W}$, is constructed. This feature map is subsequently injected into different stages of the network decoder, providing explicit geometric cues about the shape of transparent objects. It assists segmentation by sharpening curved boundaries and improves depth estimation by enhancing geometric consistency across the reconstructed surface.

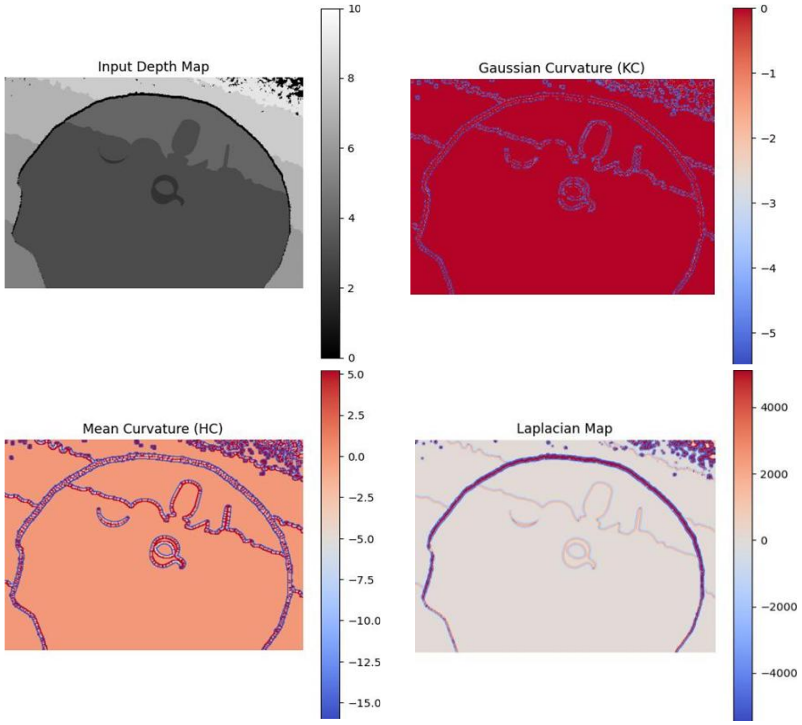


**Figure 3.** Visualization of curvature feature maps derived from depth maps

## 3.2. Curvature-Constrained Multi-Task Iterative Optimization in CESINet

Accurate depth reconstruction and semantic segmentation of transparent objects are tightly coupled tasks. Precise object contours provide crucial boundary constraints for depth estimation, while accurate depth information, in turn, helps differentiate objects from the background, thereby refining segmentation results. However, effectively leveraging this inter-task complementarity and incorporating more refined geometric priors remains a key challenge in multi-task learning. This section details how CESINet addresses these challenges by integrating curvature features into an iterative optimization process and designing a loss function that includes a curvature consistency constraint, thus jointly handling both tasks.

In CESINet's Iterative Synergistic Fusion Decoder (ISFD), the curvature features derived from the depth map, denoted as $C_{map}$, are utilized to provide explicit geometric guidance for feature fusion and optimization.

The initial curvature feature map, $C_{map} \in R^{K\times H\times W}$, can be dynamically computed from the input RGB image or from the depth map predicted in early iterations of the network. To use $C_{map}$ in different layers j of the decoder, which have varying feature map resolutions ($H_j$, $W_j$), it can be downsampled via average pooling or stride convolution to $C_{map_j} \in R^{K\times H_j\times W_j}$.

In each iteration n and at each feature scale j of the ISFD, $C_{map_j}$ can be incorporated into the feature stream as follows:

Feature Concatenation: The (potentially processed)

$C_{map_j}$ is concatenated with the depth and segmentation features fed into the SCSA module:

$$F_{d_{scsa}} = Concat(F_{d_j}^{(n-1)}, Project(C_{map_j}))$$

$$F_{s_{scsa}} = Concat(F_{s_j}^{(n-1)}, Project(C_{map_j}))$$

Here, $Project(C_{map_j})$ denotes a small, learnable projection (a 1x1 convolution) applied to $C_{map_j}$ to align its channel dimension with the task-specific features.

To further enhance the geometric realism of the prediction results, particularly for depth estimation, we introduce a curvature consistency loss term, $L_{curv}$, into the hybrid loss function. The baseline loss function is:

$$L_{hybrid} = \alpha L_{geo} + \beta L_{sem}$$

Where $L_{geo}$ encompasses the depth L2 loss, depth gradient L1 loss, and normal vector L1 loss. $L_{sem}$ is the standard cross-entropy loss. We compute the predicted curvature map $C_{pred\_map}$ from the network's predicted depth map D, and similarly, the ground truth curvature map $C_{true\_map}$ from the ground truth depth map $D^*$.

The curvature loss is then defined as the difference between these two curvature maps:

$$L_{curv} = \left|\left|C_{pred_{map}} - C_{true_{map}}\right|\right|_2$$

The final loss function for CESINet becomes:

$$L_{CESINet} = \alpha L_{geo} + \beta L_{sem} + \gamma L_{curv}$$

Here, $\gamma$ is a hyperparameter that balances the influence of the curvature constraint term. This comprehensive loss function is applied at each or the final few outputs of the ISFD's N iterations. This encourages the network to first learn coarse features and then progressively refine details, including the geometric consistency of curvature. By directly modulating curvature features within the decoder and integrating a curvature-aware loss function, CESINet is compelled to generate geometrically more plausible depth maps and semantic segmentation masks. These predictions better respect the inherent shape properties of transparent objects. The iterative optimization process allows

these curvature constraints to propagate and be reinforced across multiple processing steps, leading to synergistic improvements in both depth reconstruction and semantic segmentation tasks.

# 4. EXPERIMENTAL SETUP AND EVALUATION

## 4.1. Dataset and Experimental Platform

The experiments were conducted on the ClearPose dataset, which is specifically designed for transparent object depth perception and semantic understanding tasks. This dataset contains over 350,000 RGB-D images captured in real-world environments, providing detailed annotations for more than five million transparent object instances, including depth maps, surface normals, object categories, masks, and 6D poses. The dataset covers 63 types of transparent objects, such as common household items (e.g., glasses, bottles, and plates) as well as laboratory apparatus (e.g., test tubes and beakers).

The scenes exhibit high diversity, encompassing various challenging conditions such as indoor environments with heavy occlusions, transparent covers, mixed opaque distractors, internal fluids, and non-planar layouts. Such diversity enables a comprehensive evaluation of model robustness under realistic and complex optical conditions. The dataset is divided into training and testing subsets to assess generalization performance. The training set primarily consists of household and laboratory scenes, while the test set includes novel backgrounds, severe occlusions, and adversarial cases with opaque interference or complex layering. The schematic illustration of the ClearPose dataset is shown in Figure 4.


**Figure 4.** Representative samples from the ClearPose dataset

All experiments were implemented on a Ubuntu 22.04 LTS system equipped with an NVIDIA GeForce RTX 4090 GPU (24 GB GDDR6X memory) using CUDA 12.1 for GPU acceleration. The software environment was managed through Anaconda3, and the models were developed and trained using the PyTorch framework (Python 3.8).

The Vision Transformer (ViT-B/16) served as the primary backbone. Each training batch contained four images. The AdamW optimizer was employed due to its suitability for Transformer-based architectures. The network was trained for 150 epochs until convergence. A differential learning rate strategy was applied, with an initial rate of $1\times10^{-5}$ for fine-tuning the pre-trained ViT backbone and $3\times10^{-4}$ for the decoder to accelerate convergence. A ReduceLROnPlateau scheduler automatically reduced the learning rate when the validation loss failed to improve for five consecutive epochs. To enhance generalization, online data augmentation techniques such as random horizontal flipping, rotation, and random cropping were applied. All input images were resized to 384×384 before being fed into the network.

## 4.2. Evaluation Metrics

In the multi-task learning experiments, three standard quantitative metrics are adopted for the depth estimation task—Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Relative Error (REL)—to comprehensively measure the overall deviation, average bias, and normalized relative error, respectively. For the semantic segmentation task, Intersection over Union (IoU) and mean Average Precision (mAP) are employed, ensuring consistency with existing benchmark methods and fair comparison across models.

In the single-task experiments, the same metrics (RMSE, MAE, and REL) are used for evaluating depth estimation, while mean IoU (mIoU) and overall pixel accuracy (Acc) are adopted for segmentation. The mIoU metric reflects the overlap between the predicted segmentation mask and the ground-truth annotation, whereas Acc measures the overall pixel-level classification accuracy. Together, these metrics provide a comprehensive evaluation of the model's performance in both tasks under various scenarios.

## 4.3. Comparative Results and Analysis

To thoroughly evaluate the performance of the proposed CESINet in transparent object perception, comparative experiments were conducted from two perspectives: (1) multi-task learning frameworks and (2) single-task specialized models.

(1) Comparison with Multi-Task Methods

Representative multi-task learning frameworks, including ISGNet [18], InvPT, and TaskPrompter, were selected for comparison. These models have been widely validated on general-purpose datasets such as NYUD-v2 and represent distinct paradigms of task interaction and parameter sharing.

Experimental results on the ClearPose dataset demonstrate that all baseline methods suffer from blurry segmentation boundaries and unstable depth predictions in transparent scenes. In contrast, CESINet consistently achieves the best results across all metrics—mIoU for segmentation and RMSE for depth estimation—highlighting the advantage of the proposed curvature-enhanced prior and synergistic attention mechanism under complex optical conditions.

For fairness, the results of ISGNet were obtained directly from its official ClearPose benchmark report [18]. Since InvPT and TaskPrompter did not provide official results on ClearPose, we reproduced them under identical experimental settings as described in Section 4.1 to ensure comparability.

Qualitative results further validate this trend. As illustrated in Table 1, CESINet produces smoother and more geometrically consistent depth maps, as well as segmentation masks that align more closely with the true object boundaries, particularly in regions involving refraction and reflection.

**Table 1.** Quantitative comparison of multi-task learning methods on the ClearPose dataset

| | Model | Depth | | | Segmentation | |
|---|---|---|---|---|---|---|
| | | RMSE↓ | MAE↓ | REL | mAP | IoU |
| 1 | InvPT | 0.157 | 0.167 | 0.172 | 94.38 | 88.64 |
| 2 | TaskPrompter | 0.232 | 0.215 | 0.198 | 95.67 | 91.39 |
| 3 | ISGNet | 0.123 | 0.052 | 0.068 | 97.83 | 92.84 |
| 4 | CESINet | 0.112 | 0.050 | 0.060 | 98.27 | 94.33 |

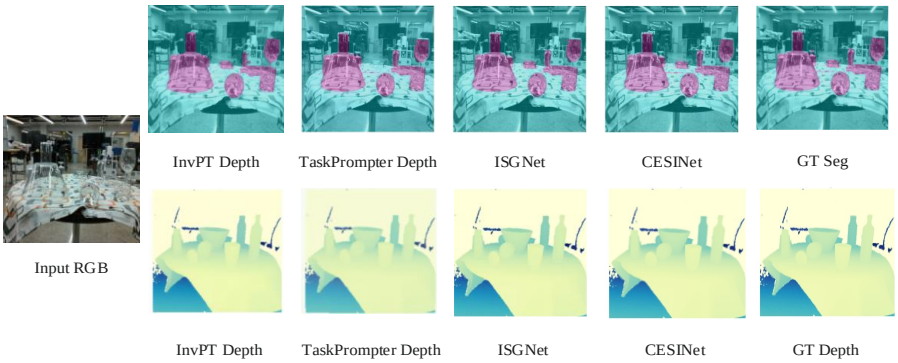


**Figure 5.** Qualitative comparison of segmentation and depth estimation results among multi-task learning methods on the ClearPose dataset

(2) Comparison with Single-Task Models

To establish an upper-bound reference for each task, CESINet was further compared with state-of-the-art single-task models designed specifically for transparent object segmentation and depth estimation.

For semantic segmentation, two representative methods were selected: DeepLabv3+ and SegFormer. The former represents a classical convolution–atrous architecture widely used in semantic segmentation, while the latter is a Transformer-based lightweight architecture that reflects recent advances in segmentation networks. Both models were retrained on ClearPose under identical conditions for fair comparison: ViT-B/16 backbone, 384×384 input resolution, batch size of 4,100 training epochs, AdamW optimizer, differential learning rates ($1\times10^{-5}$ for the backbone and $3\times10^{-4}$ for the decoder), and identical data augmentation strategies (random flipping, rotation, and cropping). Although the original DeepLabv3+ and SegFormer architectures use ResNet and MiT backbones respectively, all models here were unified with ViT-B/16 for consistent evaluation.

**Table 2.** Performance comparison of single-task segmentation models

| | Model | mIoU | ACC |
|---|---|---|---|
| 1 | DeepLabv3+ | 87.54 | 85.72 |
| 2 | SegFormer | 89.62 | 88.74 |
| 3 | CESINet | 90.21 | 89.97 |

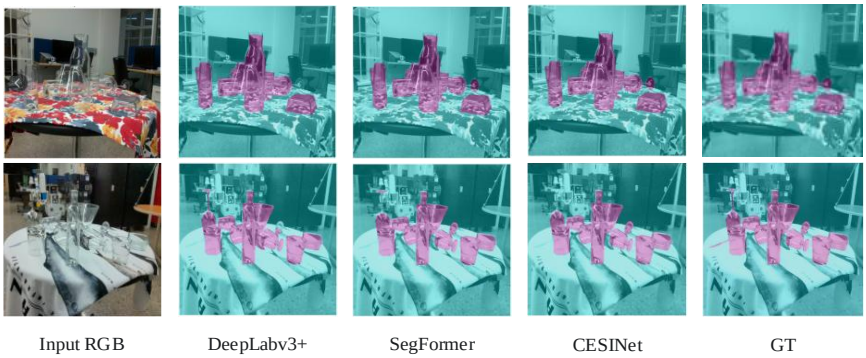


**Figure 6.** Qualitative comparison of depth estimation results among single-task models

For the depth estimation task, the official ClearPose baseline ImplicitDepth and the TransCG model—both of which report results under the ClearPose benchmark—were selected as single-task reference methods. Their results were obtained from the official evaluations reported in the ClearPose paper [5]. Since these metrics were produced under the official ClearPose evaluation protocol, which differs from our multi-task experimental setting, they are used here only as upper-bound performance references for single-task models.

**Table 3.** Performance comparison of single-task depth estimation models

| | Model | RMSE | REL | MAE |
|---|---|---|---|---|
| 1 | ImplicitDepth | 0.133 | 0.120 | 0.102 |
| 2 | TransCG | 0.077 | 0.065 | 0.060 |
| 3 | CESINet | 0.063 | 0.057 | 0.055 |

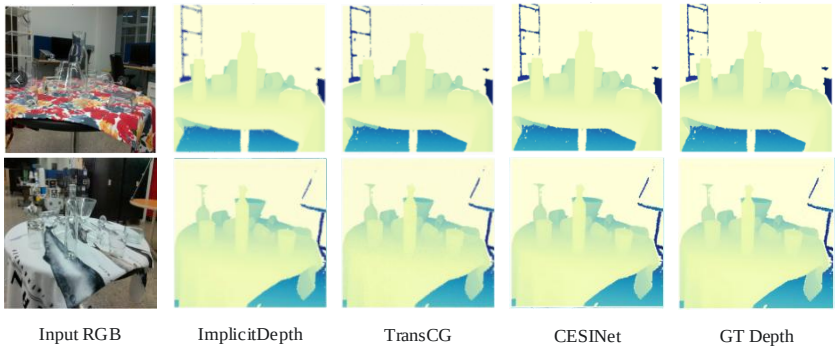


**Figure 7.** Qualitative comparison of segmentation estimation results among single-task models

The visualized results further demonstrate the predicted performance in representative transparent object scenarios. Compared with existing methods, the proposed model generates smoother and more consistent depth maps, while its segmentation boundaries align more precisely with the ground-truth contours. These observations validate the effectiveness of the proposed curvature-enhanced prior and synergistic attention mechanism in multi-task learning for transparent object perception.

### 4.4. Ablation Study

To evaluate the contribution of each component to the overall performance, an ablation study was conducted on the ClearPose test set. Specifically, three model variants were examined: removing the curvature feature (w/o Curvature Feature), removing the curvature consistency loss (w/o Curvature Loss), and removing the synergistic attention module (w/o Synergistic Attention).

As shown in Table 4, removing any of these modules leads to a simultaneous degradation in both segmentation and depth estimation metrics. In particular, when the curvature feature is removed, the IoU/mIoU scores drop significantly while RMSE and REL increase, indicating that curvature priors effectively constrain the surface geometry of transparent objects. When the curvature consistency loss is removed, the overall accuracy also declines, suggesting that this loss term reinforces the network's ability to maintain geometric coherence. Finally, the removal of the synergistic attention module weakens the mutual enhancement between the segmentation and depth estimation tasks, resulting in the most consistent yet lower performance across all metrics.

The complete CESINet configuration achieves the best results on all evaluation metrics, validating the complementarity and necessity of the three proposed components.

**Table 4.** Ablation study of CESINet components on the ClearPose dataset

| | Model | SCSA | CI | Lcurv | Depth | | Segmentation | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | RMSE↓ | MAE↓ | REL | mAP | IoU |
| 1 | Baseline | | | | 0.123 | 0.081 | 86.27 | 98.21 | 86.27 |
| 2 | Baseline+ SCSA | √ | | | 0.120 | 0.078 | 87.32 | 98.23 | 86.77 |
| 3 | Baseline+ SCSA+CI | √ | √ | | 0.110 | 0.074 | 88.13 | 98.30 | 87.23 |
| 4 | CESINet | √ | √ | √ | 0.108 | 0.072 | 88.65 | 98.33 | 87.69 |

## 5. CONCLUSION

To address the challenging problem of transparent object perception under monocular vision, this paper proposes a novel deep learning framework named CESINet. The core idea of this method departs from traditional paradigms that rely solely on the network's implicit learning of geometric features. Instead, it explicitly incorporates surface curvature, which accurately characterizes 3D shapes, as a geometric

prior directly injected into an iterative multi-task perception network. Furthermore, a three-stream synergistic attention fusion mechanism is designed to integrate this geometric prior with depth and segmentation feature flows, while being guided by a dedicated curvature consistency loss for supervision.

This work verifies the effectiveness of incorporating second-order geometric derivatives into multi-task perception. Future research may explore the feasibility of introducing higher-order or more expressive geometric descriptors as prior knowledge. Moreover, it would be promising to investigate how multi-view consistency or temporal cues from video sequences can be leveraged to learn such geometric priors in a self-supervised manner, thereby reducing the dependence on high-quality ground-truth depth data and enhancing the model's adaptability to unseen environments.

## ACKNOWLEDGMENTS

This work was supported by the Tianjin Key Research and Development Program Institute-City Cooperation Project (No.23YFYSHZ00280) and Key Natural Science Project of Tianjin Municipal Education Commission's Scientific Research Program (No. 2022ZD032, 2022ZD026).

## REFERENCES

[1] Chen G, Han K, Wong K Y K. Tom-net: Learning transparent object matting from a single image [C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 9233-9241.

[2] Sajjan S, Moore M, Pan M, et al. Clear grasp: 3d shape estimation of transparent objects for manipulation [C]//2020 IEEE international conference on robotics and automation (ICRA). IEEE, 2020: 3634-3642.

[3] Xie E, Wang W, Wang W, et al. Segmenting transparent objects in the wild [C]//European conference on computer vision. Cham: Springer International Publishing, 2020: 696-711.

[4] Kalra A, Taamazyan V, Rao S K, et al. Deep polarization cues for transparent object segmentation [C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 8602-8611.

[5] Chen X, Zhang H, Yu Z, et al. Clearpose: Large-scale transparent object dataset and benchmark [C]//European conference on computer vision. Cham: Springer Nature Switzerland, 2022: 381-396.

[6] Wang Y R, Zhao Y, Xu H, et al. Mvtrans: Multi-view perception of transparent objects [J]. arXiv preprint arXiv:2302.11683, 2023.

[7] Hamdi A, AlZahrani F, Giancola S, et al. MVTN: Learning multi-view transformations for 3D understanding [J]. International Journal of Computer Vision, 2025, 133(4): 2197-2226.

[8] Woo S, Park J, Lee J Y, et al. Cbam: Convolutional block attention module [C]//Proceedings of the European conference on computer vision (ECCV). 2018: 3-19.

[9] Lu J, Yang J, Batra D, et al. Hierarchical question-image co-attention for visual question answering [J]. Advances in neural information processing systems, 2016, 29.

[10] WANG Zhaokui, ZHOU Zhengguang, WANG Hailin, et al. MCA: Multidimensional collaborative attention in deep convolutional neural networks for image recognition [J]. Neurocomputing, 2022, 489: 497-508.

[11] Cui Y, Han C, Liu D. Cml-mots: Collaborative multi-task learning for multi-object tracking and segmentation [J]. arXiv preprint arXiv:2311.00987, 2023.

[12] Misra I, Shrivastava A, Gupta A, et al. Cross-stitch networks for multi-task learning [C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 3994-4003.

[13] Eigen D, Fergus R. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture [C]//Proceedings of the IEEE international conference on computer vision. 2015: 2650-2658.

[14] Hernández-Bautista M, Melero F J. SR-CurvANN: Advancing 3D surface reconstruction through curvature-aware neural networks [J]. Computers & Graphics, 2025: 104260.

[15] Harrison J, Benn J, Sermesant M. Improving neural network surface processing with principal curvatures [J]. Advances in Neural Information Processing Systems, 2024, 37: 122384-122405.

[16] Bhardwaj S, Vinod A, Bhattacharya S, et al. Curvature Informed Furthest Point Sampling [J]. arXiv preprint arXiv:2411.16995, 2024.

[17] da Silva S A, Geiger D, Velho L, et al. Towards Understanding 3D Vision: the Role of Gaussian Curvature [J]. arXiv preprint arXiv:2508.11825, 2025.

[18] Liu, J., Ma, H., Guo, Y., et al. (2025). Monocular depth estimation and segmentation for transparent object with iterative semantic and geometric fusion. arXiv:2502.14616.

[19] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale [J]. arXiv preprint arXiv:2010.11929, 2020.

[20] Ranftl R, Bochkovskiy A, Koltun V. Vision transformers for dense prediction [C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 12179-12188.

[21] Yang G, Tang H, Ding M, et al. Transformer-based attention networks for continuous pixel-wise prediction [C]//Proceedings of the IEEE/CVF International Conference on Computer vision. 2021: 16269-16279.

[22] Si Y, Xu H, Zhu X, et al. SCSA: Exploring the synergistic effects between spatial and channel attention [J]. Neurocomputing, 2025, 634: 129866.