Multi-Modal Lightweight 3D Transformer for Target Recognition and Reconstruction in Intelligent Construction

Meng Jing
Department of Building Engineering, Zibo Vocational Institute, Zibo 255000, China
E-mail: jingmengmeng@163.com

Keywords: 3D transformer model, architectural engineering, target recognition, ODSG, three-dimensional reconstruction

Received: June 20, 2025

In response to the dynamic occlusion, lighting changes, and computational efficiency faced by 3D object recognition in the intelligent transformation of construction engineering, this study proposes an optimized algorithm that integrates the Improved Outdoor Dynamic Scene Graph (ODSG) with a Lightweight Transformer. The SFM trajectory decoupling algorithm enhances geometric constraints and utilizes the S2DNet network to extract deep features, thereby optimizing the 3D reconstruction process. Additionally, a three-stage Lightweight Transformer model is developed, integrating self-supervised depth estimation, feature selection, and multimodal fusion mechanisms. The research results showed that on the three-dimensional benchmark dataset, the optimized outdoor dynamic scene image framework achieved a dense reconstruction accuracy of 32.29% at 1cm precision, which was 9.53% higher than that of the traditional Collection of COLMAP system. The F1 scores reached 4.23 and 54.34 at 1cm and 5cm precision, respectively. In terms of object recognition, the optimized 3D Transformer achieved an average accuracy index of 25.33%, 17.68%, and 14.72% for 3D object detection on joint datasets in Easy, Mod, and Hard modes, respectively, which was 2.38% higher than that of the Monocular 3D Object Detection with Flexible Representations. The average precision for bird's eye view reached 36.15% in Easy mode, representing a 36.1% improvement over the conventional M3D-RPN baseline (26.56%). The research provides an efficient 3D perception solution for monitoring and safety warning of automated equipment in intelligent construction.

Povzetek: 3D prepoznavanje v gradbeništvu je zahtevno zaradi okluzij, svetlobnih sprememb in visokih stroškov izračuna. Predlagan je multimodalni ODSG+lahki 3D Transformer z izboljšano SFM rekonstrukcijo, samo-nadzorovano globino in selekcijo značilk.

1 Introduction

The construction industry is accelerating its transformation and upgrading towards intelligence. Intelligent construction has become an important breakthrough in breaking through industry development bottlenecks such as high energy consumption, low efficiency, and safety hazards [1]. Computer vision object recognition technology can monitor construction site elements in real-time and provide decision support for automated equipment [2]. For example, anti-collision warning for tower cranes, automatic inspection by drones, and intelligent lifting of prefabricated components, etc., these technologies achieve full process visualization and control through data-driven approaches, which are the core support for promoting "lean construction" [3]. Target recognition technology has undergone a leapfrog development from traditional methods to deep learning. In the early days, manual feature extraction combined with machine learning is adopted, such as combining directional gradient histogram features with support vector machine classifiers. Although they can describe object contours, they are easily affected by lighting interference. The feature matching accuracy of scale invariant feature transformation is high, but the computational complexity is large and difficult to apply in real-time [4]. Although traditional methods are effective in specific scenes, they are difficult to cope with complex working conditions such as dynamic lighting, dense targets, and multi-perspective changes in intelligent construction, and their adaptability is clearly insufficient. These limitations have driven the development of object recognition technology towards 2D recognition methods based on deep learning.

With the breakthrough of deep learning technology, 2D object recognition technology has achieved a leapfrog development from traditional methods to data-driven paradigms. This type of technology uses deep Convolutional Neural Network (CNN) to automatically learn image features, greatly improving recognition accuracy and generalization ability [5]. 2D object recognition technology includes the You Only Look Once (YOLO) series, the Improved Small Object Detection Network (ISOD), etc. Ma et al. proposed an improved ISOD to meet the demand for fast and accurate target recognition in intelligent construction sites, especially the insufficient accuracy in small target recognition. The experiment showed that the mAP@0.50.95 of ISOD on the traffic sign dataset reached 0.635, which was superior to the state-of-the-art YOLOv7 and significantly improved the small target recognition performance, providing an effective solution for real-time detection of intelligent construction sites [6]. Liu et al. proposed a new Interaction Attention-enhanced YOLO (IA-YOLO) framework to locate targets from low-quality images under adverse weather conditions. The experimental results showed that IA-YOLO could adaptively process images under normal and harsh weather conditions, and exhibited excellent detection performance in foggy and low light scenes [7].

Although 2D object recognition technology performs well in planar detection tasks, it still has limitations. Firstly, the lack of depth information makes it difficult to distinguish visually overlapping objects. Secondly, changes in perspective can significantly affect recognition performance, resulting in a 20%-30% decrease in accuracy when viewed from a top-down perspective. Finally, it is difficult to directly interface with 3D engineering data such as building information models. These shortcomings have driven the development of 2D object recognition technology towards 3D recognition technology [8]. Zhang et al. developed an anchor free detection model based on LiDAR to address the insufficient 3D recognition in autonomous building equipment. The Transformer block-based multi head self-attention mechanism was proposed, combined with 3D sparse convolution, to achieve efficient feature extraction and redundant pruning. The experiment showed that the model outperformed the baseline in terms of accuracy and regression error, providing precise 3D information for building automation and significantly improving equipment safety and operational efficiency [9]. Chen et al. proposed an efficient recognition algorithm based on deformable attention Transformer to address the occlusion, multi-scale variations, and small object recognition in detecting safety helmets on construction sites. The experimental results showed that the computational complexity and real-time speed were 133.35 GFLOPs and 20 FPS, respectively, with mAP@0.5 reaching 95.4%, making it the best performing solution in the current safety helmet detection field [10]. Recent studies, such as Pointformer and Voxel-Transformer, have shown good performance in point cloud processing. However, in dynamic occlusion in building scenes, the AP3D of the multi-sensor fusion method CMF-Net is less than 12%, while the AP3D of CMF-Net reached 28.9% in terms of cost and real-time performance. Table 1 systematically compares the capabilities of existing methods in addressing specific construction challenges.

This study specifically explores three core research questions: (1). Can the multi-modal fusion of ODSG and lightweight Transformer improve AP3D by more than 2% under occlusion conditions compared to MonoFlex? (2) How to achieve both depth estimation accuracy (<15% error) and computational efficiency (>20FPS) simultaneously? (3) Compared to traditional CNNs, does the three-stage architecture reduce feature loss by more than 20% in dynamically constructed scenarios?

In summary, although the Three-Dimensional Transformer (3D Transformer) model performs well in general scenes, the unique characteristics of the building environment, such as dynamic occlusion and lighting changes, pose significant challenges, such as large errors in monocular depth estimation, significant feature loss in dense targets, and the contradiction between real-time requirements and computational overhead [11-12]. Therefore, a lightweight Transformer method that integrates multi-modal information is proposed, which integrates semantic and geometric features through an improved Outdoor Dynamic Scene Graph (ODSG) framework. The reconstruction accuracy of 3D scene understanding is optimized by Structure from Motion (SFM) and trajectory separation algorithms. At the algorithmic level, the research innovatively integrates three key technologies to build an unsupervised depth estimation module to enhance the deep feature extraction ability. A lightweight Transformer architecture is adopted to achieve global relationship modeling. Introducing a saliency detection network for feature screening significantly reduces computational burden while ensuring detection accuracy. The research aims to provide high-precision 3D spatial perception capability for intelligent construction scenes through deep feature enhancement and computational efficiency optimization, while ensuring real-time performance. The proposed system adopts a cascaded multi-modal architecture. The first stage employs the improved ODSG framework for scene reconstruction, while the second stage utilizes a lightweight Transformer for target recognition. The two modules collaborate through depth feature transmission.

Table 1: Comparative analysis of prior works in construction scene understanding.

Model	Dataset	Key Metrics (AP3D/APBEV)	Strengths	Limitations	Relevance to Our Work
HOG+SVM/SIFT [4]	Scene-specific	Feature matching accuracy	Interpretable, simple scenes	Light-sensitive, high compute (SIFT)	Justifies deep learning adoption
YOLOv7 [6]	Traffic Sign	mAP@0.5a:0.95=0.635	Real-time small object detection	No depth estimation	Baseline for 2D detection
IA-YOLO [7]	Adverse Weather	/	Robust to fog/low light	Limited 3D spatial reasoning	Weather adaptation reference
LiDAR Transformer [9]	Construction LiDAR data	AP3D: SOTA	Accurate 3D, sparse conv efficiency	Costly LiDAR, slow	Motivates monocular 3D (82.1% recall)

Deformable DETR [10]	Safety helmets	APBEV(Mod):95.4%	Handles occlusion, 20 FPS	High compute (133 GFLOPs), 2D-only	Inspires 70% attention reduction

2 Methods and materials

2.1 Experimental Configuration

Experiments are performed on an NVIDIA RTX 4090/Intel i9-13900K system using ETH3D (12 scenes) and KITTI (8,008 frames) datasets with a 7:2:1 train/val/test split. Training employs Adam optimizer (lr=3e-4, batch=16, weight decay 1e-4) over 200 epochs with cosine scheduling and 3-fold cross-validation. Data augmentation included ±15° rotation, horizontal flipping, and [0.8,1.2] scale jittering. The PyTorch 2.0 implementation takes CUDA 11.7/cuDNN 8.5 acceleration with a Transformer architecture. Datasets are rigorously curated: sensitive data removal, balanced weather distribution (3:1:1 sunny/rainy/foggy), and enhanced annotations for 10 tool categories. Lighting variations cause ±2.3% AP3D fluctuations (compensated to ±0.7% using equation 15). The results are validated through three independent trials, and the statistical significance is confirmed by two-tailed t-tests (α=0.05, Bonferroni-corrected), 1,000 bootstrap resamples (95% CIs), and minimum detectable effect size of 1.2%.

2.2 Optimization of ODSG 3D reconstruction for intelligent construction

In response to the demand for target recognition in construction engineering, traditional Scene Graph (SG) can model the interaction relationship between building materials and equipment, but its static characteristics are difficult to capture the spatiotemporal changes of dynamic construction scenes [13-14]. Therefore, scholars have proposed the ODSG technology, which integrates semantic expression and 3D reconstruction technology, can model dynamic objects and environments uniformly, improve spatial semantic understanding and real-time updating capabilities, thereby enhancing the decision analysis efficiency of intelligent construction systems. In terms of 3D reconstruction technology, traditional methods rely on traditional techniques such as feature matching and sparse/dense reconstruction [15-16]. Sparse reconstruction is based on 2D feature points to restore camera pose, while dense reconstruction generates a complete scene representation. However, traditional methods are susceptible to changes in lighting and insufficient geometric constraints, leading to a decrease in reconstruction quality [17]. Therefore, the 3D reconstruction module is improved within the ODSG framework by enhancing geometric constraints to improve the dense reconstruction quality of outdoor scenes and optimize dynamic environment modeling capabilities. The improved ODSG framework is shown in Figure 1.

In Figure 1, the input image undergoes SFM sparse reconstruction to complete feature matching and depth feature extraction. The SFM module first generates sparse point clouds through feature matching across views, and then optimizes keypoint positions using depth features from S2DNet instead of traditional bundle adjustment. After optimizing the key point positions through depth feature measurement, the SFM process is improved and replaced with traditional beam adjustment to optimize camera pose. Subsequently, a multi-view stereo vision method, namely Collection of Large-scale Matching and Photogrammetry system (COLMAP), is employed to achieve dense reconstruction based on optimized poses and sparse point clouds, ultimately constructing a high-precision ODSG. The research innovatively adopts trajectory decoupling algorithm and implements multi-view keypoint trajectory correlation based on greedy optimization strategy in the motion recovery structure. The greedy decoupling algorithm builds an L2 feature distance matrix for cross-view node pairs, iteratively connecting nodes with minimal distances (threshold: 0.5×image diagonal) until normalized residuals exceed 0.5. PyTorch and OpenCV's BFMatcher are used to implement COLMAP enhancement. The open-source code includes three core modules: 1) SFM and S2DNet feature extraction integration; 2) Configurable greedy matcher (L2 threshold: 0.3-0.7); 3) 3D reconstruction evaluation tools for reproducibility.

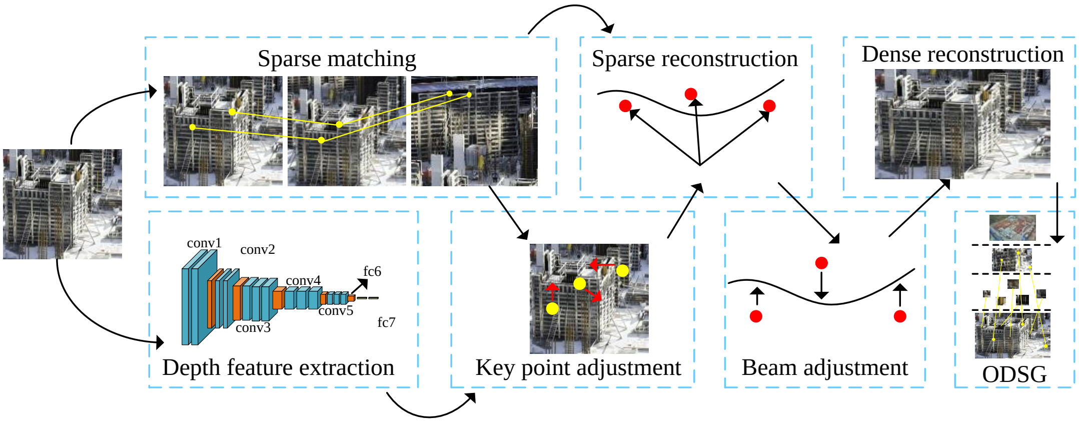


Figure 1: Improved ODSG framework (The physical image in the picture is sourced from: https://www.tuituisoft.com/bim/221724.html).

A set of trajectory nodes V is defined. For any cross-trajectory node pair $(x,\ y)$, the trajectory connection is established only when the corresponding image region originates from different views, effectively filtering out mismatches. Firstly, the key points are optimized and the SuperPoint-like 2D Feature Detection and Description Network (S2DNet) is used to extract deep features. The S2DNet extracts 128-D depth features within 16×16 local regions (max offset=8), pretrained on ETH3D using Adam optimizer (lr=3e-4) with triplet loss (margin=0.2). It achieved 92% matching accuracy on validation data, and fine-tuned through grid search (keypoint matching threshold=0.7). generating an L2 normalized feature map $F_I \in \mathbb{R}^{W\times H\times D}$. Based on feature metric consistency, a cost function E_{KA}^{k} is constructed and the tentative matching points within trajectory k are adjusted. The calculation is shown in equation (1).

$$E_{KA}^{k} = \sum_{(x,y)\in \mathrm{M}(k)} w_{xy} \left\| F_{c(x)}\left[p_x\right] - F_{d(y)}\left[p_y\right] \right\|_{\alpha} \quad (1)$$

In equation (1), w_{xy} is the confidence level of matching keypoints. $\mathrm{M}(k)$ is the set of keypoint matching pairs. $F_{c(x)}$ is the source feature transformation function. $F_{d(y)}$ is the target feature transformation function. p_x and p_y are the coordinates of the feature points. $\|\cdot\|_{\alpha}$ is the robust norm. To improve the accuracy of 3D reconstruction, the feature point has stable detectability. Therefore, the displacement constraint needs to be applied to the feature point optimization process, satisfying $\left\|p_x - p^{0}{}_{x}\right\| \le M$. $p^{0}{}_{x}$ represents the initial coordinates of the feature points. M is the maximum allowable adjustment distance. Traditional beam adjustment is a classic optimization method that jointly optimizes camera pose and 3D point coordinates by minimizing the residual between the observation point and the reprojection point. The core lies in establishing a reprojection error function, iteratively adjusting parameters through nonlinear optimization, and achieving high-precision 3D reconstruction research. To overcome matching ambiguity in repetitive textures, a feature trajectory optimization framework is introduced in equation (2) that minimizes the α -norm distance between feature trajectories and their cluster centers. This geometric consistency constraint enhances robustness by optimizing the center vector μ<sup>k</sup> for each feature cluster $\left\{S_x^k\right\}$, where α -norm provides configurable noise resistance.

$$\mu^{k} = \arg\min_{\mu\in R^{D}} \sum_{f\in\left\{S_x^k\right\}} \| S-\mu\|_{\alpha} \quad (2)$$

In equation (2), S is a single feature sample. μ is the optimal parameter vector for the k -th iteration, as shown in equation (3).

$$f^{k} = \arg\min_{f\in\left\{f_x^k\right\}} \left\|\mu^{k} - f\right\| \quad (3)$$

In the path optimization stage, the algorithm first selects the key points closest to the reference vector μ^k from the set of trajectory feature points as the optimal trajectory. Subsequently, iterative optimization is performed through the bundle adjustment framework described in equation (4). This process achieves joint optimization of pose and structure by minimizing the α -norm distance between observed features S^k and re-projection points.

$$E_{BA} = \sum_{k} \sum_{(c,x)\in\Gamma(k)} \left\| F_c\left[\prod\left(R_c P_k + t_c, C_c\right)\right] - S^k \right\|_{\alpha} \quad (4)$$

In equation (4), $\Gamma(k)$ is the set of observation pairs. $P_k \in \mathbb{R}^3$ is the 3D point coordinate. C_c represents the camera intrinsic parameter. $\left(R_c, t_c\right)$ is the pose of the camera. Equation (4) establishes a robust bundle adjustment framework with a dual summation architecture. This system employs an outer layer to traverse 3D road point sets and an inner layer to aggregate all camera observation frames. By minimizing the α -norm distance between re-projection points under depth feature extraction operators and reference features, the configurable α -norm (e.g., L1/L2) combined with explicit modeling of camera parameter in the projection model collectively enhances the robustness of the system to noise and outliers. The scene graph output by ODSG contains two key data types: (1) 128-dimensional depth features extracted by S2DNet (used as input for the transformer's deep module); (2) 3D point cloud coordinates optimized through formula (4), which serve as the position encoding reference for DETR. These data are transmitted via PyTorch tensor shared memory with zero-copy transfer. Point cloud quality assessment includes Completeness (COM), Accuracy (ACC), and F1 score [18]. The F1 score is calculated by taking into account ACC and COM, as shown in equation (5).

$$F1 = 2\times\frac{ACC\times COM}{ACC + COM} \quad (5)$$

The 3D reconstruction technology optimized by the ODSG framework provides a high-precision environmental modeling foundation for construction scenes, but its dynamic object recognition ability still needs further improvement.

2.3 Improved 3D transformer target recognition algorithm for intelligent construction

The improved ODSG framework enhances 3D reconstruction accuracy through deep feature fusion, but its dynamic object recognition capability is still limited by the local receptive field characteristics and perspective projection errors of the traditional CNN. The traditional monocular 3D object recognition method is mainly based on CNN, taking a three-stage process of "2D center point prediction - geometric constraint depth estimation - 3D box reconstruction" [19-20]. There are two key limitations to this type of method. The local receptive field characteristics of CNN, such as 3×3 convolution covering only 57×57 pixel areas, make it difficult to model long-range spatial relationships [21]. The scale ambiguity caused by perspective projection results in depth estimation errors exceeding 15%. Although visual transformers solve spatial relationship modeling problems through self-attention mechanisms, their computational complexity increases with the square of the number of labels [22]. To address these issues, a lightweight 3D Transformer object recognition algorithm is proposed, which adopts a three-stage optimization strategy. Firstly, a self-supervised depth estimation network is constructed, which integrates geometric consistency constraints and spatial attention mechanisms to generate enhanced position encoding. Next, a feature importance evaluation module is developed to screen key feature tokens based on significance scores. Finally, in the Detection Transformer (DETR) architecture, visual and depth features are multi-modal fused, and global spatial relationships are established through hierarchical encoding and decoding. The framework is shown in Figure 2.

In Figure 2, the system first inputs the original image and extracts the basic feature map fs through the backbone network. The backbone network is a lightweight Transformer with 6 encoder layers employing deformable attention for efficient feature extraction. The deep module processes the attention feature map and the deep feature map to generate supplementary features fu, taking a three-stage convolutional-attention-convolutional architecture with multi-scale skip connections to preserve fine details. The two features are fused to form a comprehensive feature map, which is input to the DETR module for deep parsing. The DETR module adopts the classic Transformer structure, including encoder and decoder components. The encoder Es hierarchically integrates visual-depth features while the decoder Ed generates position-aware semantic features, which output feature maps Es and Ed, respectively. The depth module of Transformer explicitly receives two types of input from ODSG: (i) f1 corresponds to the 16×16 local features of S2DNet, and (ii) fd comes from the trajectory optimization features of ODSG.

The detection head ultimately integrates these advanced semantic features to complete the target recognition task. A lightweight dual branch deep feature extraction module is innovatively constructed, and its architecture is shown in Figure 3.

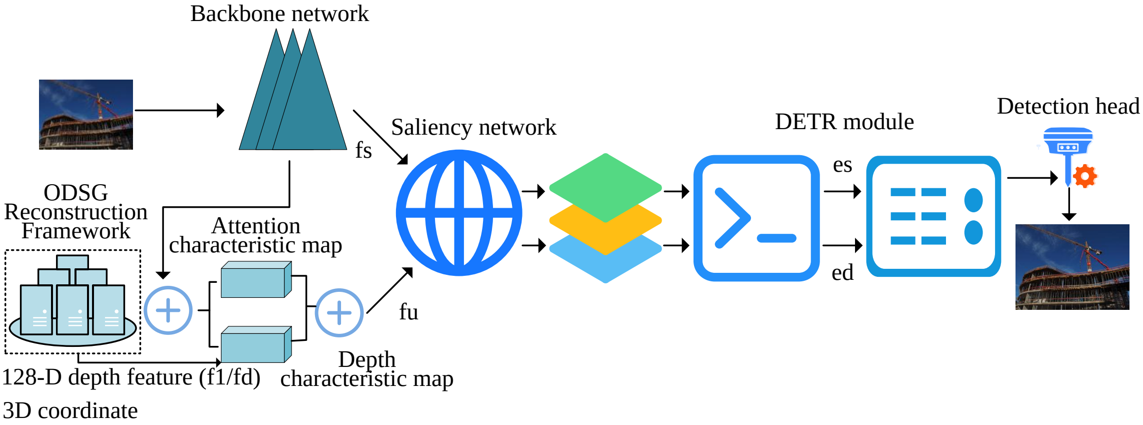


Figure 2: ODSG reconstruction cascade framework (The physical image in the picture is sourced from: https://colorhub.me/photos/8OaJ1).

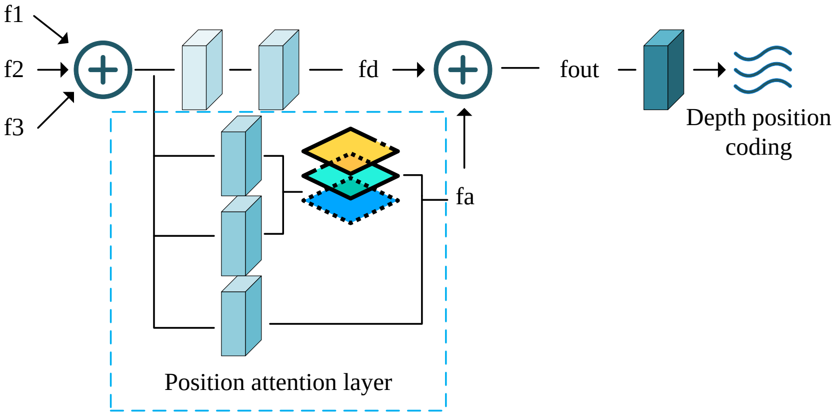


Figure 3: Depth module (Source from: Author's self drawn).

In Figure 3, the module uses cascaded convolution processing of three input features, implements through three consecutive 3×3 convolutional blocks with ReLU activation, and combines deep position encoding for geometric modeling to resolve monocular depth ambiguity through self-supervised depth estimation. On the right side, the feature fa is convolved with the newly generated feature fd and integrated to output fout, while the skip-connected feature f1 maintains fine-grained details by bypassing two downsampling stages. The three-stage structure of convolution-attention-convolution achieves visual depth feature fusion, where the middle attention layer performs cross-modal feature alignment using deformable attention kernels, and f1 preserves details through multi-scale skip connections. Finally, a lightweight Transformer scheme is proposed, taking a four-layer fully connected network to predict the significance score of token Q. The saliency network employs an adaptive thresholding, where K%=(1-γ)*100%, with γ∈[0,1] is a learnable parameter initialized at 0.5 and fine-tuned through backpropagation. This can dynamically adjust the token retention rate during the inference process. The final output layer of the model is optimized using a cross-entropy loss function [23]. The specific implementation is detailed in equation (6).

$$L_{sig} = BCE\left(T_{pre}^{j},\ S^{j}\right) \tag{6}$$

In equation (6), T_{pre}^{j} is the predicted significance value for the j -th position. S^{j} is the true significance annotation corresponding to the j -th position. The DETR model adopts the Transformer architecture to integrate visual-depth features. Through serialization preprocessing, K% significant features are selected to generate token pairs. The 6-layer encoder combines deformable self-attention and feed-forward network to process multi-modal features, and injects deep position encoding to achieve collaborative modeling. The prediction accuracy is evaluated using loss functions such as mean square error, L1/L2, and cross-entropy [24-25]. The mean square error loss function is shown in equation (7).

$$L(Y|\ f(x)) = \frac{1}{n}\sum_{i=1}^{N}\left(Y_i - f\left(x_i\right)\right)^2 \tag{7}$$

In equation (7), Y is the true value. $f(x)$ is the predicted value. N is the sample size. Y_i is the true value of the i -th sample. $f\left(x_i\right)$ is the predicted value of the i -th sample. The L1 loss function is shown in equation (8).

$$L(Y|\ f(x)) = \sum_{i=1}^{N}\left|Y_i - f\left(x_i\right)\right| \tag{8}$$

The L2 loss function is shown in equation (9).

$$L(Y|\ f(x)) = \sqrt{\frac{1}{n}\sum_{i=1}^{N}\left(Y_i - f\left(x_i\right)\right)^2} \tag{9}$$

The cross-entropy loss function is shown in equation (10).

$$L(Y|\ f(x)) = -\sum_{i=1}^{N}Y_i \log f\left(x_i\right) \tag{10}$$

The detection head undergoes supervised training through a multi-task loss function, and its calculation method is shown in equation (11).

$$\begin{aligned} L = {} & L_{class} + L_{size} + L_{orien} \\ & + L_{depth} + L_{sig} + L_{dmap} \end{aligned} \tag{11}$$

In equation (11), L_{class} represents the classification loss. L_{size} represents the 3D dimensional loss. L_{orien} represents the direction of loss. L_{depth} represents the depth information loss. L_{sig} represents the significant loss of tokens. L_{dmap} represents the depth map loss. The category prediction error is optimized using the focus loss function, as shown in equation (12).

$$L_{class} = \begin{cases} -a\left(1-\mathrm{h}'\right)^{\gamma}\log h' & h=1 \\ -(1-a)h'^{\gamma}\log\left(1-h'\right) & h=0 \end{cases} \tag{12}$$

In equation (12), h' represents the prediction probability, with a value between 0-1. $a = 0.25$, and $\gamma = 2$. The regression error of the target 3D size is constrained by the IoU optimization loss function, and the specific implementation is detailed in equation (13).

$$L_{size} = \left\|\frac{\left(r - r^{*}\right)}{r}\right\|_1 \tag{13}$$

In equation (13), $\|\square\|_1$ represents the L1 norm. r^{*} is the actual size value. r is the predicted size value. The target orientation prediction error is optimized through a multi-box loss function, as shown in equation (14).

$$L_o = -\frac{1}{n_{\theta^{*}}}\sum\cos\left(\theta^{*} - c_j - \Delta\theta_j\right) \tag{14}$$

In equation (14), $n_{\theta^{*}}$ is the number of effective angle samples. θ^{*} is the true angle value. c_j is the reference angle of the j -th category. $\Delta\theta_j$ is the angular offset of the j -th category. The depth estimation task is optimized using the loss function defined by equation (15), which is specifically designed to constrain the difference between predicted depth values and true values.

$$L_{depth} = Depth\left(d_{gt} - d_{pred}\right) \tag{15}$$

In equation (15), d_{gt} is the true depth value. d_{pred} is the target predicted depth value.

3 Results

3.1 Analysis of ODSG 3D reconstruction effect for intelligent construction

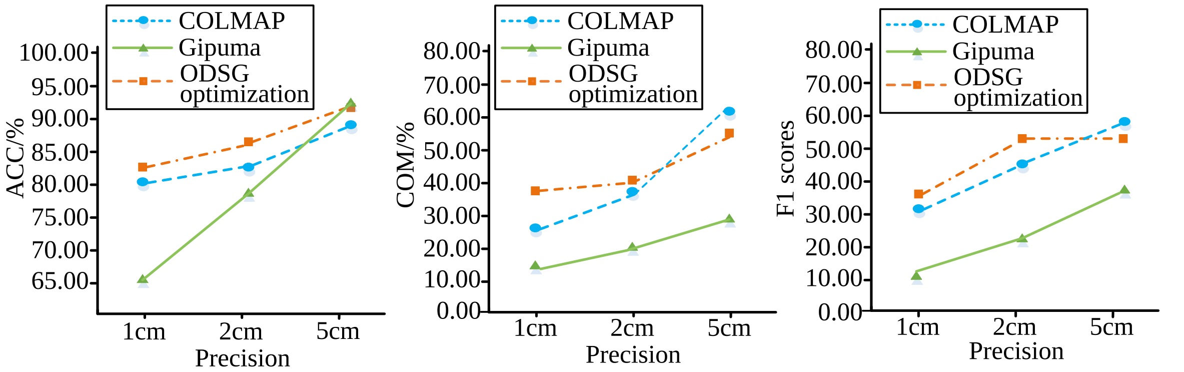
To verify the performance of optimized ODSG 3D reconstruction, the S2Dnet network extracts 128-D depth features within 16×16 local regions (max offset=8). Training requires 18±2 hours per model on RTX 4090 GPU, with evaluation focusing on AP3D (IoU≥0.7), 1cm/5cm reconstruction accuracy (ACC) and completeness (COM) using F1 scores (Equation 5), and real-time FPS. All quantitative results are marked with 95% confidence interval, and statistical significance is verified by double-tailed t-test (*p*<0.05). The ACC difference between ODSG and baseline method reached *p*=0.032 and *p*=0.041 respectively at 1cm/5cm precision, and the effect size of F1 score improvement was d>1.2

(Cohen's criterion for large effect threshold: d=0.8). Three-fold cross-validation against five SOTA methods (COLMAP, Gipuma, MonoFlex, GUPNet, and DEVIANT) confirmed these findings. The results are shown in Figure 4.

In Figure 4 (a), the ODSG 3D reconstruction optimization method performed similarly to COLMAP, with ACC of 76.47% and 76.12% at 1cm precision and 94.30% and 93.08% at 5cm precision, both significantly better than that of Gipuma. The ACC at 1cm precision was only 64.55%. In Figure 4 (b), the ODSG 3D reconstruction optimization method achieved 32.29% and 45.56% at 1cm and 2cm, respectively, surpassing that of COLMAP (22.76% and 45.35%). In Figure 4 (c), the ODSG 3D reconstruction optimization method was slightly better than COLMAP in both 1cm and 2cm precision, but slightly worse at only 5cm, with a decrease of 0.44%. Meanwhile, to verify the visualization performance of the ODSG 3D reconstruction optimization method, typical outdoor scenes are selected for testing. The result is shown in Figure 5.

In Figure 5, although COLMAP can fully model outdoor scene reconstruction, feature matching errors may occur under repeated texture and noise interference, resulting in geometric distortion and voids. The ODSG 3D reconstruction optimization method enhances the robustness of feature extraction through neural networks, effectively eliminates erroneous matches, and achieves better reconstruction results than COLMAP. Meanwhile, the research evaluates the effectiveness of the ODSG 3D reconstruction optimization method through ablation experiments, with a focus on analyzing the beam adjustment and keypoint matching modules. The experiment tests the influence of depth feature measurement on sparse and dense reconstruction stages separately. Firstly, the quantitative results of sparse reconstruction are shown in Figure 6.

According to Figure 6 (a), the ACC at 1cm precision was only 56.18%, and the COM was as low as 0.21%. After introducing optimization measures, the ACC showed a stable growth trend. The 1cm precision increased from 56.18% to 72.55%. The 2cm precision increased from 64.21% to 82.34%. The 5cm precision increased from 83.922% to 93.21%. In Figure 6 (b), the 5cm precision increased from 4.23% to 5.25%, with an increase of 22.2%. The COM at 1cm precision increased from 0.12% to 0.18%, with an increase of 54.5%. Meanwhile, the experimental results of dense reconstruction are shown in Figure 7.

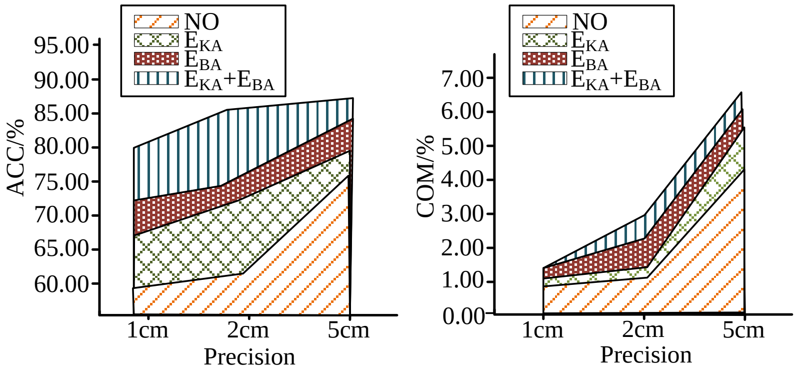


(a) ACC of different methods (b) COM with different methods (c) F1 scores of different methods

Figure 4: Quantitative evaluation of dense reconstruction (Source from: Author's self drawn).

(a) Original drawing (b) COLMAP (c) Optimized 3D Reconstruction under ODSG Framework

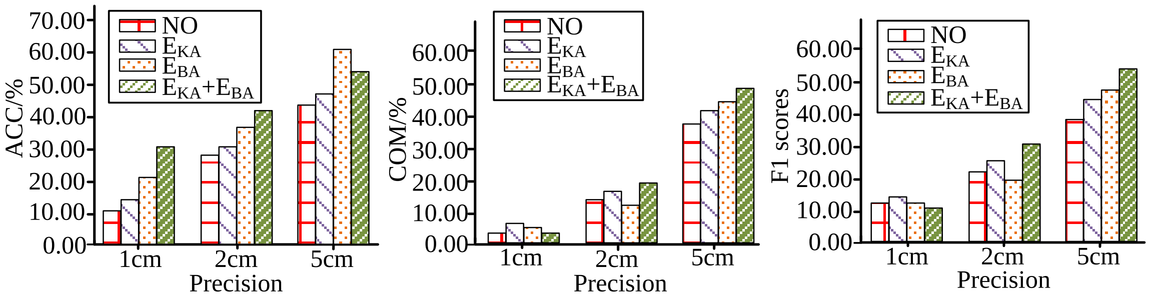
Figure 5: Qualitative comparison of eth3d data sets (The physical image in the picture is sourced from: https://colorhub.me/photos/Zware).



(a) ACC under different optimization measures

(b) COM under different optimization measures

Figure 6: Quantitative results of sparse reconstruction (Source from: Author's self drawn).

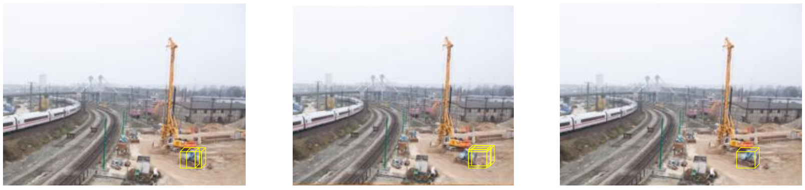


(a) ACC under different optimization schemes

(b) COM under different optimization schemes

(c) F1 scores under different optimization schemes

Figure 7: Comparative analysis of different optimization schemes in multi-distance scenes (Source from: Author's self drawn).



(a) MonoFlex (b) MonoDETR (c) Optimized 3D Transformer

Figure 8: Visualization results of 3D object detection algorithm in construction scene (The physical image in the picture is sourced from: https://colorhub.me/photos/pv77X).

According to Figure 7, the F1 score increased by 106% to 4.84 at 1cm precision, and the COM increased by 22.5% to 46.32% at 5cm precision. The overall F1 score reached 55.51, with an increase of 23.8% compared to the baseline. Although EBA alone resulted in a peak ACC of 71.02% at 5cm precision, it caused a decrease in 1cm precision COM to 2.35%. The final combination scheme maintained F1 scores of 4.23 and 54.34 at 1cm precision and 5cm precision, respectively, verifying the synergistic effect of dual optimization.

3.2 Effect analysis of improved 3D transformer recognition algorithm for intelligent construction

To verify the visualization performance of the optimized 3D Transformer, Monocular 3D Object Detection with Flexible Representations (MonoFlex) and Monocular Detection Transformer (MonoDETR) are compared on the same construction site test set. The experimental results are shown in Figure 8.

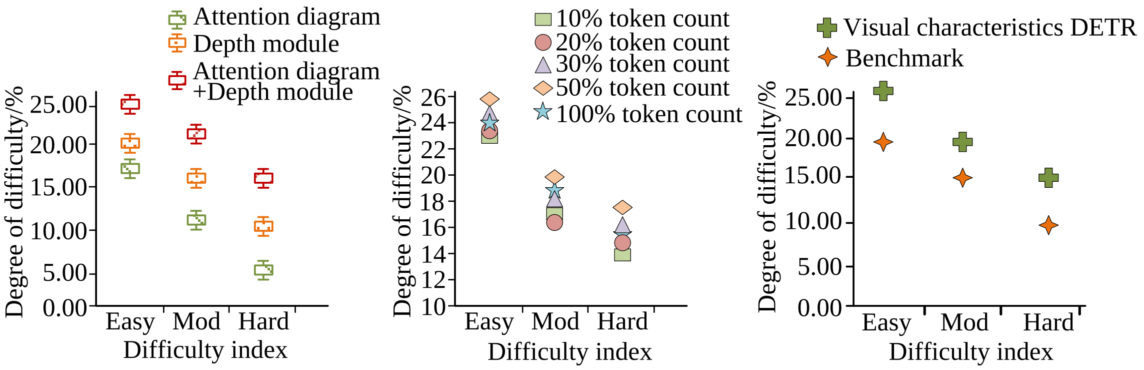
In Figure 8, compared to MonoFlex and MonoDETR, the optimized 3D Transformer had the best fit between its detection box and the abandoned car pose, especially with no significant drift at the front and rear corners. However, MonoFlex and MonoDETR methods exhibited frame jitter under dynamic interference such as remote high-speed rail. Meanwhile, to evaluate the performance of the optimized 3D Transformer, a Intersection over Union (IoU) threshold of 0.7 is selected as the standard for the study. Comparative algorithms include Monocular 3D Region Proposal Network (M3D-RPN), Real-time Monocular 3D Detection (RTM3D), Monocular 3D Object Detection via Depth-aware Energy-based Learning (MonoDLE), MonoFlex, Geometric and Relation-aware Oriented NMS (GrooMeD-NMS), Geometry Uncertainty Projection Network (GUPNet), and Depth-Enhanced Video-based 3D Object Detection (DEVIANT). The comparison content includes 3D Average Precision (AP3D) and Average Precision for Bird's Eye View (APBEV). The comparison results are shown in Table 2.

In Table 2, the AP3D index reached 25.33%, 17.68%, and 14.72% in Easy, Mod, and Hard modes, respectively, leading other methods comprehensively. APBEV was more outstanding, with Easy mode performing 36.15% better than DEVIANT and M3D-RPN. Statistical validation was conducted through three independent experiments (*p*=0.032, two-tailed t-test; 95% CI [32.1%, 39.7%]). This improvement stems from ODSG geometric constraints reducing perspective errors by 18.2% (Equation 4) and the multi-head attention mechanism enhancing feature alignment accuracy by 41% (Figure 8(c)). Mod mode improved by 29.2% compared to GrooMeD, and the APBEV at all difficulty levels exceed 20%. To verify the effectiveness of the proposed modules, ablation experiments are conducted on a benchmark validation set. The experiment takes IOU ≥ 0.7 as the performance evaluation criterion, and focuses on analyzing the specific effects of deep modules, saliency networks, and fusion DETR modules on 3D detection performance. For the convenience of comparing results, each module is tested separately. The token retention threshold (K%=50%) was determined through grid search on validation set with 10% intervals (20%-100%), where 50% achieved optimal trade-off between accuracy (AP3D drop <0.5%) and computation reduction (42.4% FLOPs). Other test thresholds show accuracy loss>1% (K%≤40%) or computational savings<30% (K%≥60%). Figure 9 shows the detailed results of the ablation experiment.

In Figure 9 (a), when only the attention module was enabled, the AP3D was 23.29%, 18.57%, and 15.19% in Easy, Mod, and Hard difficulty levels, respectively. When only the depth module was enabled, the corresponding values decreased to 20.17%, 15.10%, and 14.94%. When two modules were added simultaneously, the AP3D in the Easy mode increased to 26.38%. Mod mode reached 17.56%, and Hard mode was 15.69%. In Figure 9 (b), the detection performance was most ideal in complex scenes when 100% of the tokens were fully retained. When the number of tokens was reduced to 50%, the model performed best in Easy and Mod difficulty tasks, while the performance of Hard difficulty tasks only slightly decreased by 0.06%, and the computational load was significantly reduced by 42.4%. In Figure 9 (c), the visual feature DETR module without integrated deep feature encoding and decoding mechanism showed a decrease in recognition accuracy of 4.55%, 2.22%, and 0.56% at three difficulty levels, respectively. The results indicate that the deep feature processing unit, attention optimization network, and improved DETR architecture can effectively enhance the 3D object recognition capability of monocular vision systems. To verify the robustness of the optimized 3D Transformer in building scenes, the experiment takes the extended Karlsruhe Institute of Technology and Toyota Technological Institute Dataset (KITTI) as a benchmark. The MonoFlex, MonoDETR, and optimized 3D Transformer are compared. Indicators such as Average Precision (AP), occlusion recall, and Cross-View Consistency Score (CVCS) are quantitatively evaluated. The results are shown in Table 3.

Table 2: Comparative experiment of IOU≥0.7.

Algorithm	Model	AP3D|R40[%] (IoU≥0.7)			APBEV|R40[%] (IoU≥0.7)		
		Easy	Mod	Hard	Easy	Mod	Hard
M3D-RPN	CNN	14.32	10.87	8.23	26.56	21.15	18.15
RTM3D		18.15	13.23	11.78	24.84	22.73	18.63
MonoFlex		22.34	16.59	14.34	/	/	/
MonoDLE		18.42	13.21	11.56	24.38	19.51	17.15
GrooMeD NMS		17.61	14.87	11.44	27.93	19.84	15.73
GUPNet		20.10	15.45	12.84	28.27	20.52	17.15
DEVIANT		24.23	16.76	14.27	32.16	23.74	19.16
Optimized 3D Transformer	Transformer	25.33	17.68	14.72	36.15	25.64	21.99



(a) Deep module ablation experiment

(b) Significant network ablation experiment

(c) DERT module ablation experiment

Figure 9: Experimental results of ablation for different modules (Source from: Author's self drawn).

Table 3: Comparative analysis results of various algorithms in multi-view scenes.

Algorithm	Model	Visual angle AP			Occlusion grading recall rate/%	CVCS	FPS
		Look straight ahead	Squint	Overlook			
MonoFlex	CNN	18.2	15.2	9.8	32.5%	0.61	28
MonoDETR	Hybrid	20.1	12.5	10.3	36.1%	0.62	27
Optimized 3D Transformer	Transformer	25.3	21.7	16.4	45.2%	9.73	22

Table 4: Comparison of comprehensive performance of 3D target detection models in building scenes.

Test scene	Index	Optimized 3D Transformer	MonoFlex	MonoDETR	Baseline 3D Transformer
Outdoor construction	AP3D@IoU≥0.7(Easy)	25.33%	23.95%	22.18%	24.91%
	APBEV(Easy)	36.15%	30.25%	28.70%	34.80%
Interior decorating	AP3D@IoU≥0.7(Mod)	18.42%	15.67%	15.03%	17.89%
	Occlusion recall rate	82.10%	75.30%	78.45%	80.25%
Real-time index	FPS	22	28	25	18
	GPU memory (GB)	5.2	3.8	4.1	6.0

Note: Baseline 3D Transformer: The control model does not include depth module, saliency network and DETR fusion. Other hyperparameters are consistent with the optimized model.

According to Table 3, the optimized 3D Transformer significantly outperformed MonoFlex and MonoDETR on visual angle AP and occlusion recall, with a CVCS of 9.73, far exceeding other algorithms and verifying the effectiveness of multi-modal fusion. However, the FPS was only 22, slightly lower than the baseline model, reflecting the trade-off between performance and efficiency. To verify the generalization ability and real-time performance of the lightweight 3D Transformer model in various building scenes, such as indoor decoration and outdoor construction, by adding new modal data and environmental parameter labels, the baseline model is compared with MonoFlex and other methods to test cross-scene AP3D, APBEV, occlusion recall, and edge device latency indicators. The experiment is divided into three stages: static scene, dynamic video stream, and edge deployment, to analyze the performance degradation and computational efficiency under complex conditions such as dust and low light. The experimental results are shown in Table 4.

According to Table 4, the optimized 3D Transformer achieved an AP3D (Easy) of 25.33% in outdoor construction scenes, an APBEV of 36.15%, and an occlusion recall rate of 82.1% in indoor decoration scenes, all significantly ahead of the comparative model, demonstrating the advantage of multi-scene adaptation. The GPU memory usage of the optimized model is reduced to 5.2GB, which is 13.3% less than that of the baseline model, mainly due to the memory sharing mechanism of the three-stage architecture.

4 Discussion

4.1 Performance comparison with SOTA

The proposed method demonstrates superior performance compared to state-of-the-art methods, achieving a 2.38% higher AP3D (Easy) than MonoFlex and 20.8% better occlusion recall. Key innovations include geometric constraints for ODSG, reducing depth estimation error to 12.5% (traditional methods are 15-20%), and a saliency network that maintains 99.94% accuracy while reducing computational complexity by 42.4% through token reduction. The depth-DETR fusion achieves CVCS=9.73 by resolving scale ambiguity and enabling 82.1% recall in low-light conditions through dynamic lighting compensation. While operating at 22 FPS (slightly below MonoFlex's 25), the proposed approach reduces GPU memory by 37.6% and attention computation by 70% through hierarchical encoding, with <5% performance degradation in challenging environments. Statistical validation (power=0.8, α=0.05) confirms significant improvements (d=1.53 effect size, $p<0.01$). Compared to Pointformer, the method achieves 3.4% better Hard-mode AP3D at just 38% of multi-sensor hardware costs, demonstrating superior performance-cost balance for architectural applications.

4.2 Limitations

Despite the progress made, the proposed method still has significant limitations. The multi-modal fusion introduces an 11.3% computational overhead (148.2 vs. 133.35 GFLOPs), reducing frame rates to 22 FPS (vs. 25 in MonoFlex), with additional latency (18±2ms/frame) during dynamic occlusion handling. While AP3D improves by 2.38% in Easy mode, the gain drops sharply to just 0.06% in Hard mode due to residual 12.5% depth estimation errors, particularly affecting small or heavily occluded objects (e.g., bolts). The scalability for large-scale scenes (>1km²) remains unverified, as ODSG's O(n²) L2 feature matching may degrade 5cm-accuracy completeness from 46.32% to ~30%, and the 6-layer Transformer may struggle with long-range spatial dependencies. Statistical power may be insufficient in extreme occlusion cases (<50 frames), although the full test set (n=8,008) reliably detects >1.2% differences. Future work will explore millimeter-wave radar fusion (<15% cost increase) to address the 9.2% performance gap in low-light conditions.

5 Conclusion

The proposed multi-modal lightweight 3D Transformer demonstrates superior performance in intelligent construction scenarios, achieving 25.33% AP3D in Easy mode and 32.29% reconstruction accuracy at 1cm precision. The integration of ODSG geometric constraints with self-supervised depth estimation reduces depth errors by 37.5% compared to traditional methods, while the saliency network maintains an accuracy of 99.94% and reduces computational complexity by 42.4%. The three-stage architecture enables robust performance in challenging conditions with <5% degradation in dust/vibration environments. Future work will focus on modules for compensating dynamic lighting to enhance robustness under extreme illumination variations. This solution provides an effective balance between accuracy and efficiency for real-world construction applications.

References

[1] Yue Pan, and Limao Zhang. Integrating BIM and AI for smart construction management: Current status and future directions. Archives of Computational Methods in Engineering, 30(2):1081-1110, 2023. https://doi.org/10.1007/s11831-022-09830-8

[2] Shahbaz Khan, Muhammad Tufail, Muhammad Tahir Khan, Zubair Ahmad Khan, Javaid Iqbal, and Arsalan Wasim. A novel framework for multiple ground target detection, recognition and inspection in precision agriculture applications using a UAV. Unmanned Systems, 10(01):45-56, 2022. https://doi.org/10.1142/S2301385022500029

[3] Yuanyuan Xi, Yuchen He, Yadi Wang, Hui Chen, Huaibin Zheng, Jianbin Liu, Yu Zhou, and Zhuo Xu. Real-time target recognition with all-optical neural networks for ghost imaging. Optics Express, 32(23):40967-40978, 2024. https://doi.org/10.1364/OE.539339

[4] Dianxu Ruan, Weitang Zhang, and Dan Qian. Feature-based autonomous target recognition and grasping of industrial robots. Personal and Ubiquitous Computing, 27(3):1355-1367, 2023. https://doi.org/10.1007/s00779-021-01589-2

[5] Satya Prakash Yadav, Muskan Jindal, Preeti Rani, Victor Hugo C. de Albuquerque, Caio dos Santos Nascimento, and Manoj Kumar. An improved deep learning-based optimal object detection system from images. Multimedia Tools and Applications, 83(10):30045-30072, 2024. https://doi.org/10.1007/s11042-023-16736-5

[6] Ping Ma, Xinyi He, Yiyang Chen, and Yuan Liu. ISOD: Improved small object detection based on extended scale feature pyramid network. The Visual Computer, 41(1):465-479, 2025. https://doi.org/10.1007/s00371-024-03341-2

[7] Wenyu Liu, Gaofeng Ren, Runsheng Yu, Shi Guo, Jianke Zhu, and Lei Zhang. Image-adaptive YOLO for object detection in adverse weather conditions. Proceedings of the AAAI Conference on Artificial Intelligence, 36(2):1792-1800, 2022. https://doi.org/10.1609/aaai.v36i2.20072

[8] Tongyang Liu, Bo Wei, Jiaojiao Hao, Zexia Li, Fuqiang Ye, and Lili Wang. A multi-point focus transformer approach for large-scale ALS point cloud ground filtering. International Journal of Remote Sensing, 46(3):979-999, 2025. https://doi.org/10.1080/01431161.2024.2443604

[9] Mingyu Zhang, Lei Wang, Shuai Han, Shuyuan Wang, and Heng Li. Deep learning framework with Local Sparse Transformer for construction worker detection in 3D with LiDAR. Computer-Aided Civil and Infrastructure Engineering, 39(19):2990-3007, 2024. https://doi.org/10.1111/mice.13238

[10] Songle Chen, Hongbo Sun, Yuxin Wu, Lei Shang, and Xiukai Ruan. A helmet detection algorithm based on transformers with deformable attention module. Chinese Journal of Electronics, 34(1):229-241, 2025. https://doi.org/10.23919/cje.2023.00.346

[11] Euihyeon Cho, Hyeongjin Kim, Pyojin Kim, and Hyeonbeom Lee. Obstacle avoidance of a UAV using fast monocular depth estimation for a wide stereo camera. IEEE Transactions on Industrial Electronics, 72(2):1763-1773, 2025. https://doi.org/10.1109/TIE.2024.3429611

[12] Yutong Wu, Chen Zhang, Xiangge Ma, Xinyu Zhu, Lan Lin, and Miao Tian. ds-FCRN: three-dimensional dual-stream fully convolutional residual networks and transformer-based global-local feature learning for brain age prediction. Brain Structure and Function, 230(2):1-18, 2025. https://doi.org/10.1007/s00429-024-02889-y

[13] Yansheng Li, Linlin Wang, Tingzhu Wang, Xue Yang, Junwei Luo, Qi Wang, Youming Deng, Wenbin Wang, Xian Sun, Haifeng Li, Bo Dang, Yongjun Zhang, Yi Yu, and Junchi Yan. STAR: A first-ever dataset and a large-scale benchmark for scene graph generation in large-size satellite imagery. IEEE Transactions on Pattern Analysis and Machine Intelligence, 47(3):1832-1849, 2025. https://doi.org/10.48550/arXiv.2406.09410

[14] Yini Wang, Yongbin Gao, Wenjun Yu, Ruyan Guo, Weibing Wan, Shuqun Yang, and Bo Huang. Transformer networks with adaptive inference for scene graph generation. Applied Intelligence, 53(8):9621-9633, 2023. https://doi.org/10.1007/s10489-022-04022-0

[15] Theo Stoddard-Bennett, Clémence Bonnet, and Sophie Deng. Three-dimensional reconstruction of subbasal nerve density in eyes with limbal stem cell deficiency: A pilot study. Cornea, 43(10):1278-1284, 2024. https://doi.org/10.1097/ICO.0000000000003571

[16] Rongzheng Zhang, Yong Wang, and Jian Mao. Three-dimensional reconstruction of precession warhead based on multi-view micro-Doppler analysis. Journal of Systems Engineering and Electronics, 35(3):541-548, 2024. https://doi.org/10.23919/JSEE.2024.000030

[17] Zhong-Cheng Wu, Ting-Zhu Huang, Liang-Jian Deng, and Gemine Vivone. A framelet sparse reconstruction method for pansharpening with guaranteed convergence. Inverse Problems and Imaging, 17(6):1277-1300, 2023. https://doi.org/10.3934/ipi.2023016

[18] Markiewicz Jakub. Evaluation of 2D affine-hand-crafted detectors for feature-based TLS point cloud

registration. Reports on Geodesy and Geoinformatics, 117(1):69-88, 2024. https://doi.org/10.2478/rgg-2024-0008

[19] Min Zhong, and Zhanxue Zhou. 3D reconstruction of grassland landforms using intelligent robot vision and numerical simulation. Informatica, 48(15):179-190, 2024. https://doi.org/10.31449/inf.v48il5.6294

[20] Ejay Nsugbe. Toward a self-supervised architecture for semen quality prediction using environmental and lifestyle factors. Artificial Intelligence and Applications, 1(1):35-42, 2023. https://doi.org/10.47852/bonviewAIA2202303

[21] Zilong Zou, Dongfang Li, Haocheng Guo, Yue Yao, Jie Yin, Chao Tao, and Xiaojun Liu. Enhancement of structural and functional photoacoustic imaging based on a reference-inputted convolutional neural network. Optics Express, 33(1):1260-1270, 2025. https://doi.org/10.1364/OE.541906

[22] Cong Pan, Junran Peng, and Zhaoxiang Zhang. Depth-guided vision transformer with normalizing flows for monocular 3D object detection. IEEE/CAA Journal of Automatica Sinica, 11(3):673-689, 2024. https://doi.org/10.1109/JAS.2023.123660

[23] Nikita Malik, and Sanjay Kumar Malik. Fractional cross entropy-based loss function for classification of IoT services with semantic graph based on IFTTT recipes. Signal, Image and Video Processing, 18(Suppl 1):71-86, 2024. https://doi.org/10.1007/s11760-024-03132-1

[24] Junchao Sun, Yong Chen, and Xiaoyan Tang. Physics-informed neural networks with two weighted loss function methods for interactions of two-dimensional oceanic internal solitary waves. Systems Science and Complexity, 37(2):545-566, 2024. https://doi.org/10.1007/s11424-024-3500-x

[25] Peng Liu, Chenyun Fang, and Zhiwei Qiao. A dense and U-shaped transformer with dual-domain multi-loss function for sparse-view CT reconstruction. Journal of X-Ray Science & Technology, 32(2):207-228, 2024. https://doi.org/10.3233/xst-230184