Computers & Industrial Engineering 207 (2025) 111325

Contents lists available at ScienceDirect

Computers & Industrial Engineering

journal homepage: www.elsevier.com/locate/caie

MFMDepth: MetaFormer-based monocular metric depth estimation for distance measurement in ports☆

Xinqiang Chen [a,b], Fei Ma [a], Yuzheng Wu [a,c], Bing Han [d,e], Lijuan Luo [f,*], Salvatore Antonio Biancardo [g]

[a] Institute of Logistics Science and Engineering, Shanghai Maritime University, Shanghai 201306, China
[b] Chongqing Key Laboratory of Green Logistics Intelligent Technology, Chongqing Jiaotong University, Chongqing 400074, China
[c] SPG Qingdao Port Group Co., Ltd., Qingdao, China
[d] Shanghai Ship and Shipping Research Institute Co., Ltd., Shanghai 200135, China
[e] College of Physics and Electronic Information Engineering, Minjiang University, Fujian 350108, China
[f] Key Laboratory of Brain-Machine Intelligence for Information Behavior (Ministry of Education and Shanghai), School of Business and Management, Shanghai International Studies University, Shanghai 201620, China
[g] Department of Civil, Construction and Environmental Engineering (DICEA), University of Naples Federico II, 80125, Italy

ARTICLE INFO

Keywords:
Automated port
Automated guided vehicle
Metric depth estimation
Relative depth estimation
Transformer

ABSTRACT

Automated container guided vehicle (AGV) navigation systems that rely on magnetic pin infrastructure are currently facing serious sustainability challenges. The monocular metric depth estimation method is emerging as a promising alternative by aligning the predicted depth with the actual scale, however, its deployment in dynamic port environments (e.g., line of sight occlusion, texture loss) is still limited by precision and reliability tradeoffs. The proposed framework is implemented in three steps. Firstly, the framework initially introduces an innovative MetaFormer architecture, which incorporates a global squeeze block (GSB). This GSB employs a Squeeze Former (SF) to facilitate comprehensive modeling of inter-token relationships across the global image context. Secondly, the bins module, which combines wavelet transform convolution (WBM), is utilised to estimate the metric depth, and the backbone network is employed to estimate the relative depth. Finally, the framework fuses the two depths to achieve a refined metric depth estimation. Extensive evaluation shows that our method achieves approximately 26.3% RMSE performance improvement compared to MiDas, approximately 21.3% performance improvement on AbeRel metrics compared to SOTA model ZoeDepth. Compared to traditional baseline DS-SIDE, our approach achieves approximately two times improvement in depth prediction accuracy across all metrics, while maintaining competitive inference performance.

1. Introduction

The evolution of intelligent transportation systems (ITS) in modern ports is pivotal to sustaining global trade efficiency, with maritime logistics handling over 80 % of worldwide goods circulation (Raza et al. 2023, Chen, Huang, et al. 2025). As a cornerstone of port ITS, Automated Guided Vehicles (AGVs) for container transportation directly govern operational safety, cost-effectiveness and throughput capacity (Huang et al. 2025). However, prevailing AGV navigation systems relying on magnetic nail infrastructure face critical sustainability challenges, continuous vehicular compression degrades magnetic markers at a rate exceeding 12 % annually in high-traffic terminals, necessitating biweekly recalibrations that incur maintenance costs up to $245,000 per kilometer annually (Aizat et al. 2023). This deterioration induces cumulative localization errors exceeding ±15 cm, jeopardizing collision avoidance in congested port environments and constraining scalability across heterogeneous terminal layouts (Reis et al. 2023). More importantly, the traditional system lacks the ability to perceive the environment, and it is difficult to cope with the dynamic scenarios of the port (e.g., temporary storage yard changes, moving obstacles, etc.), which has become a key bottleneck restricting the upgrade of the smart port (Chen, Cheng, et al. 2023, Chen, Piao, et al., 2023, Chen, Hu, et al. 2025). While vision-based navigation presents a promising infrastructure-free alternative, its deployment in dynamic port settings (occluded sightlines and

☆ This article is part of a special issue entitled: 'AI-based Approaches' published in Computers & Industrial Engineering.

* Corresponding author.
E-mail address: luolijuan@shisu.edu.cn (L. Luo).

https://doi.org/10.1016/j.cie.2025.111325

Available online 18 June 2025

repetitive visual patterns) remains fundamentally constrained by the precision-reliability trade-off in monocular depth estimation (MDE).

Monocular depth estimation, which reconstructs 3D geometry from 2D images, has emerged as a critical enabler for autonomous navigation (Arampatzakis et al., 2023). Current MDE paradigms bifurcate into relative depth estimation (capturing ordinal relationships) and metric depth estimation (recovering physically scaled distances) (Xiaogang et al. 2020). Relative depth estimation typically uses visual cues between objects in an image (e.g., changes in size with distance, perspective relationships, occlusion relationships) to approximate the depth of individual pixels. For the relative depth estimation paradigm, there is no need for in-camera and out-of-camera parameters used to calibrate the true image depth, enabling generalization across scenes (Mertan et al. 2022). However, the unknown scale factor between different frames, the lack of actual physical significance of the estimated depth values (Zhao et al. 2020, Poggi et al., 2021), make this paradigm a fatal limitation for AGV path planning that requires centimeter-level accuracy (Chen, Wu, et al., 2023). Conversely, metric depth estimation aligns predictions with real-world dimensions through supervised learning (Sun et al., 2021), yet faces three unresolved challenges: (1) Boundary distortion and global consistency dilemma. Convolutional encoders in metric networks exhibit limited receptive fields, causing edge artifacts that misguide AGV steering decisions (Wu et al. 2020). While vision transformers (ViTs) mitigate this via self-attention, their quadratic complexity $O(n^2)$ escalates computational costs by 3–5 times compared to CNNs, rendering real-time deployment impractical (Zhang et al. 2024). (2) Scale ambiguity in multi-scenario adaptation. Existing metric models trained on generic datasets (e.g., KITTI) suffer catastrophic performance drops (e.g., RMSE $>$ 4 of DS-SIDE in Table III) when deployed in novel port environments, as they fail to disentangle scene-specific scale factors from intrinsic geometry. (3) Underutilized hierarchical features. Current frameworks treat relative and metric depth estimation as mutually exclusive tasks, discarding the rich spatial priors embedded in relative depth predictions—an oversight shown to degrade fine-grained accuracy by from 13.7 % to 21 % in cluttered scenes (Bhat et al. 2023).

To bridge these gaps, we propose a MetaFormer–based monocular metric depth estimation framework for port AGV visual navigation, a hybrid framework synergizing metric-scale recovery with multi-frequency geometric analysis. Our work advances AGV visual navigation through three system innovations:

1) We innovatively propose a monocular metric depth estimation framework based on the Metaformer architecture and Bins module, which integrates feature information from both relative depth and metric depth. Compared to existing methods, this combination enables more effective capture of multi-frequency features in the scene and global cues between objects, achieving fine-grained metric depth estimation across different port scenarios.
2) Based on the traditional Transformer architecture, we propose the global squeeze block(GSB), which simultaneously compresses the channel dimensions of the Query and Key matrices in attention computation through Squeeze Former(SF) and adds an adaptive spatial weighting map. This effectively controls computational resource consumption while maintaining excellent global information relationship modeling capabilities.
3) We propose a novel integrated wavelet transform convolution bins module(WBM) that implements a refined metric depth estimation. This module significantly expands the receptive field of traditional conv2d while maintaining parameter efficiency. Compared to existing metric depth module designs, this structure can capture multi-scale feature frequency and spatial location information, enhancing the ability of model to recognize object shapes rather than traditional texture features.

2. Related works

2.1. Monocular metric depth estimation

Monocular metric depth estimation has emerged as a critical research direction in computer vision due to its broad applications in autonomous driving, robotics, and augmented reality. Recent advancements can be categorized into three paradigms: supervised and unsupervised learning with geometric constraints.

The Supervised method relies on precisely labeled training data that provides accurate depth labels for all pixels of each RGB image in the data set. Models trained using this paradigm can learn precise mapping relationships from pixels to depth values, but acquiring large depth-labeled datasets is expensive. Ren et al. propose a two-stage framework that first trains a classifier to distinguish between low and high depth range images, and then trains different depth estimation networks for two different depth ranges, similar to our proposed idea of fusing relative depth and measuring depth (Ren et al. 2019). Guizilini et al. realized the metric depth estimation of images by introducing variational latent features and modeling each point depth prediction as a probability distribution estimation during the network decoding phase (Guizilini et al. 2020). Bhat et al. revolutionized deep regression by expressing it as an ordered classification problem through adaptive classification. Their Bins module uses learnable bin centers and Laplacian losses to discrete depth ranges. They achieved SOTA performance on NYU depth V2 and the design of our proposed WBM is also inspired by this idea (Bhat et al. 2021, Bhat et al. 2022). Subsequent work by Hu et al. introduced the lifelong learning (LL) model, which achieved a skilled handling of the characteristics of different domains by building an indeterminate LL solution, and led the benchmarks by from 8 % to 15 % (Hu et al. 2023). Hu et al. propose Metric3d V2, a geometric foundation model that achieves zero sample generalization depth estimation performance, which proposes a canonical camera space conversion module for resolving metric ambiguity in various camera models and large-scale datasets, and can be easily integrated into existing monocultural depth estimation models (Hu et al. 2024).

Unsupervised learning with geometric constraints directly learns the inherent structure and information features of an input unlabeled data set. This kind of method usually uses the input continuous image sequence to restore the geometry of the known scene or the parallax of the same object in successive frames to create training constraints. However, while this paradigm alleviates the need for manual annotation of large-scale data, it often fails in the face of dynamic objects or low-texture (or repeated texture) regions, because dynamic objects are often treated as anomalies in unsupervised frames and their motion is coupled to camera motion, resulting in inaccurate depth estimation. Zhang et al. pioneered an inertial visual fusion framework that uses inertial measurement unit (IMU) luminosity loss and cross sensor luminosity consistency loss to align temporal IMU tracks with spatial visual features, enabling scale sensing depth prediction with a 3.5 % reduction in absolute relative error (AbsRel) compared to a pure visual baseline. However, this approach requires synchronous imu camera calibration, limiting its applicability in consumer-grade devices (Zhang et al. 2022). In response to camera parameter variations, Yin et al. developed a canonical camera space transformation module that learns internally perceived depth scaling through several new geometric consistency losses. Their approach is at the leading edge of 5 zero-shot benchmarks, albeit at the cost of increased memory consumption due to camera parameter embedding (Yin et al. 2023). Liu et al. propose an unsupervised learning framework to improve the long-term tracking stability of simultaneous localization and mapping(SLAM) systems by learning forward and backward inertial sequences on multiple word Spaces and fusing monocular depth estimation results with inertial measurement results (Liu et al. 2023). Yu et al. studied the problem of unsupervised depth estimation of images under inclined viewing angles and applied it to the field of view of unmanned aerial vehicles. They

propose an unsupervised perfect learning system consisting of a multi-resolution feature fusion depth network and a perception-improving network to obtain higher-quality depth estimation results by matching parallax results from different angles in a complex environment (Yu et al. 2023). Zhu et al. address the problem that current unsupervised methods do not effectively utilize the rich temporal features in continuous RGB image sequences. They designed a temporal optical flow mask that effectively identifies and excludes the static pixels between adjacent frames, thus improving the computational efficiency of the network (Zhu et al. 2024).

2.2. MetaFormer-based architecture

Recent research shows that the evolution of vision transformers has shifted from a purely attention-based model to a hybrid architecture that balances precision and efficiency. Below we analyze three key axes of development, namely Token mixer innovation, lightweight design, and attention myopia.

Metaformer is a generic architecture based on the Transformer structure, with its core being the flexible selection of token mixer (Yu et al. 2022). The original vision transformer (ViT) proposed by Dosovitskiy et al. demonstrates the global acceptance domain through multi-head self-attention (MHSA), which is the first time the Transformer structure has been applied to the image domain, but with quadratic complexity relative to image resolution (Dosovitskiy et al., 2020). Subsequent research explored the use of a space-reduced attention layer (SRA) to replace the traditional MHSA as a new token mixer, with a 30 % reduction in the number of parameters compared to the then-popular ResNeXt (Wang et al. 2021). The Swin Transformer proposed by Liu et al. uses shift window attention to reduce computing load (Liu et al. 2021). Yar et al. proposed a new tokenization mechanism based on spatial feature variation, which provides a larger viewable field for the network (Yar et al. 2024). Our proposed GSB architecture builds on these foundations, combining spatial weight adaptive mechanism and novel token mixer SF.

For edge deployment, Thwal et al. developed OnDev-LCT using deep separable convolution and channel transformation. Experiments on the CIFAR10 dataset show that OnDev-LCT reduces the number of parameters by 83 % compared to ViT, and the classification accuracy increases by about 25 % (Thwal et al. 2024). Mohammed et al. proposed a lightweight visual transformer based on model distillation, which is trained using semi-supervised learning based on pseudo-labels, and learns discriminative representations from labelled and unlabelled data, thereby obtaining a lightweight model (Mohammed et al. 2024). It is worth noting that the EdgeFormer proposed by Luo et al. introduced a sparse edge-sensing pixel selector and a multi-scale effective transformer module to realize adaptive computation of different image regions (Luo et al. 2024).

Venkataramanan et al. proposed the SKIPAT method, which utilizes the self-attention computation from previous layers to approximate the attention in one or more subsequent layers, significantly improving model throughput without increasing computational resources (Venkataramanan et al. 2023). Complementary work by Raptakis et al. developed Fourier domain attention using FFT-based token mixing, reducing the frequency recognition error RMSE for noisy speech input by 29.95 % (Raptakis & Pantazis, 2024). Our wavelet convolution metric module extends this concept with a learnable frequency filter that prioritizes mission-critical spectral components. In recent work, Huang et al. propose a new channel attention approximation algorithm that reduces the attention computation to linear complexity and uses ReLu to exclude irrelevant weights from the attention diagram (Huang et al. 2024). Finally, we put together a comparison Table 1 of our work and existing methods, in order to better show the differences between different methods.

Table 1
Comparison with Existing Works.

Method	Core Innovation	Metric Depth Handing	Key Limitation	AbsRel ↓
Ren et al., 2019	Range-specific networks	Dual-regime depth prediction	Discontinuous depth transitions	0.113
Bhat et al., 2021	Adaptive bin classification	Discretized depth bins	Fixed bin ranges	0.098
Zhang et al., 2022	IMU-visual fusion	Scale-aware via sensor data	Requires hardware synchronization	0.108
Hu et al., 2024	Camera space transformation	Joint depth-normal optimization	Huge amount of computation	0.067
Ours.	Wavelet-guided metric regression	Unified metric-relative fusion	Requires wavelet pre-training	0.046

3. Method

3.1. Overview

The proposed monocular metric depth estimation framework for assisting port AGV visual navigation (referred to as MFMDepth) is shown in Fig. 1. The framework consists of two main steps: Encoder and Decoder. Firstly, in the Encoder stage, we employ the GSB to modeling the global information relationships of the input image tokens, capturing the complex dependencies between pixels in the input sequence. Then, in the Decoder stage, we simultaneously predict two branches: relative depth and metric depth. The relative depth features are directly predicted by the backbone network, represented by the solid lines in Fig. 1. The metric depth features are predicted by the WBM branch, denoted by the dashed lines in Fig. 1. Finally, the relative depth features and metric depth features are added together and fed into specific heads to obtain the metric depth estimation or segmentation results. MFMDepth provides an innovative monocular metric depth estimation method that combines relative and metric depth information, achieving accurate and efficient prediction performance. In the following sections, we will expand on the design details of each component structure in this framework, from the processing flow of image tokens to the final prediction results. Finally, we elucidate our training strategy and loss function.

3.2. Encoder

The Encoder consists of a backbone network composed of orderly stacked GSBs, which achieves long-distance dependency modeling of the input sequence, thereby obtaining a coherent and consistent depth feature representation. The GSB is a generic architecture based on the Metaformer (as shown in Fig. 2), which employs an improved attention mechanism called Squeeze Former as the token mixer. By adopting a spatial weight adaptive adjustment strategy and compressing the matrix computation dimensions, it enhances model performance while reducing the computational burden. This innovation enhances the attention accuracy in port scenarios where background noise and sparse textures are prevalent.

Since the standard Transformer structure only accepts 1-dimensional token embeddings sequences as input, in the initial stage of our proposed model, we reshape a single RGB image I into a series of flattened 2-dimensional patches I_p, as shown in Eq. (1).

$$I_p = reshape(I), I \in \mathbb{R}^{H\times W\times C}, I_p \in \mathbb{R}^{N\times\left(P^2\times C\right)} \tag{1}$$

where (H, W) represents the resolution size of the RGB image, and C is the number of channels in the RGB image. N is the number of generated patches, calculated as $N = HW/P^2$. (P, P) is the size of each 2-dimen-

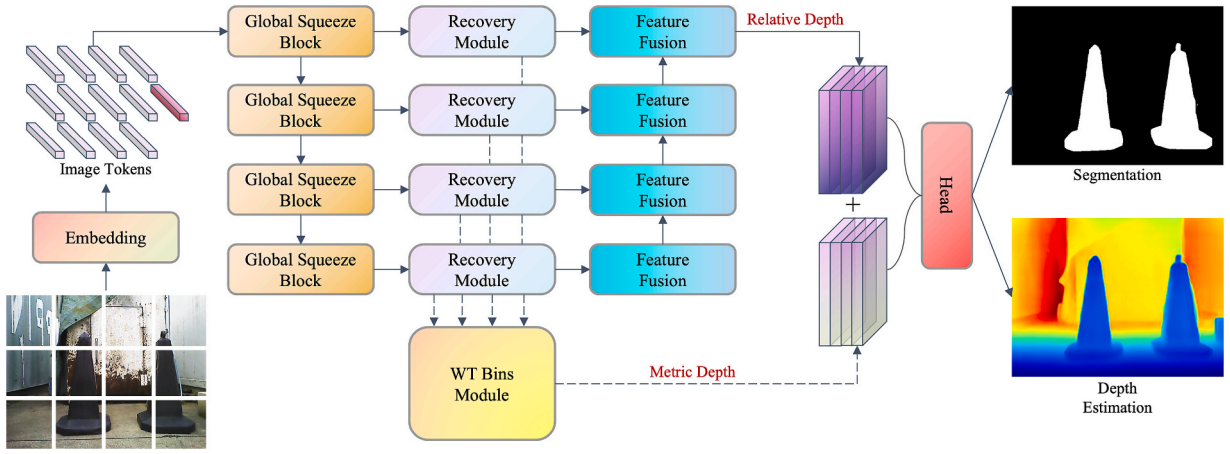

Fig. 1. General framework of MFMDepth. The encoder is composed of multiple GSB stacks, and the decoder consists of two branches. The solid line is the relative depth prediction branch, and the dashed line is the metric depth prediction branch.

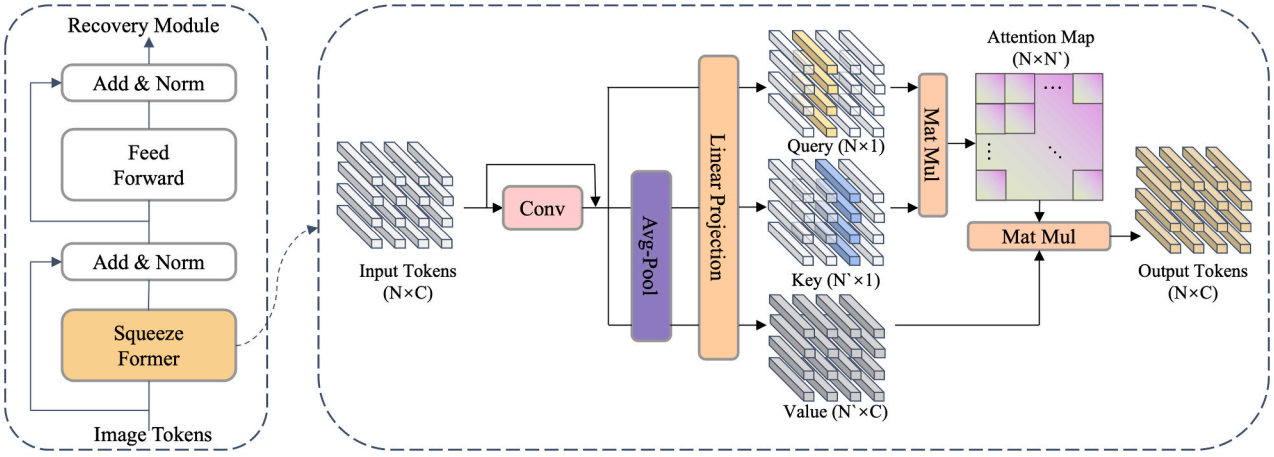

Fig. 2. Overall architecture of GSB. The dashed box on the right shows the structure of the SF.

sional patch. From a high-level perspective, these individually embedded image patches play the role of tokens in the feature space. Finally, we use a trainable linear projection layer to uniformly project each generated image patch to dimension $D = 1024$. For dense prediction tasks like depth estimation, the spatial position of pixels is crucial for predicting the depth details of object edges. Therefore, we add learnable position embeddings E_{pos} to the image embeddings to compensate for the information loss caused by reshaping the image I. Lastly, following the work in NLP, we add a learnable class-token I_{class} at the beginning of the image tokens, serving as the final global image representation used for classification. The mathematical representation of this step is shown in Eq. (2).

$$F_0 = \left[I_{class}; I_p^1 E; I_p^2 E; \cdots; I_p^N E;\right] + E_{pos}, E \in \mathbb{R}^{\left(P^2 \times C\right) \times D}, E_{pos} \in \mathbb{R}^{(N+1) \times D} \tag{2}$$

where F_0 represents the input image tokens in Fig. 2, E is the identity matrix used to supplement the dimensions. In the left dotted box of Fig. 2, the data flow of the entire GSB module is shown. As in Eq. (3–4), the image tokens first pass through the SF module, then undergo normalization and residual connection, and finally replace the SF with a feedforward neural network to repeat the appeal operation again.

$$M_l = LN(SF(F_{l-1})) + F_{l-1}, l = 1 \cdots L \tag{3}$$

$$Z_l = LN(\mathrm{FFD}(M_l)) + M_l, l = 1 \cdots L \tag{4}$$

where M_l represents the intermediate variable, Z_l represents the high-dimensional tokens that will be fed into Recovery Module. The novel token mixer structure SF, simultaneously considers the importance of global feature extraction and the computational efficiency of the self-attention mechanism for dense prediction tasks. Before the attention operation, in order to compensate for the possible loss of spatial discriminant ability in the process of channel compression, we first use convolution to extract the spatial information of input feature $F_{l-1} \in \mathbb{R}^{(N+1) \times D}$, generate a preliminary weight map, and then normalize it to ensure a reasonable weight distribution. The formula is defined as Eq. (5).

$$W_{sp} = \sigma(Conv(F_{l-1})) \tag{5}$$

where W_{sp} is the weight map, $Conv$ represents the use of standard convolution layer to extract spatial features, σ is a normalization function (e.g., sigmoid or using softmax in spatial dimensions) to map the result to the range [0, 1], giving the weight for each spatial position. In traditional self-attention, for each attentional head, we compute the query-Q, key-K and Value-V. The obtaining process is shown in Eq. (6).

$$Q = F_{l-1} W_j^Q, K = AvgPool(F_{l-1}) W_j^K, V = AvgPool(F_{l-1}) W_j^V \tag{6}$$

where $AvgPool$ represents the average pooling layer, W_j^Q, W_j^K, $W_j^V \in \mathbb{R}^{D \times d_k}$ are projection matrices, d_k is the dimension projected into the subspace, j denotes the index of the attention head. In order to reduce the computational complexity, SF compresses Q and K by channel

dimension. Define the compressed mapping function $P = \mathbb{R}^{d_k} \rightarrow \mathbb{R}$ (implemented by learnable parameters), then there is the compression process mathematical expression Eq. (7). After compression, Q' and K' are both one-dimensional representations, which significantly reduces the computational effort.

$$Q' = P(Q) \in \mathbb{R}^{(N+1)\times 1}, K' = P(K) \in \mathbb{R}^{(N+1)\times 1} \tag{7}$$

where Q' and K' represents the compressed attention matrix. Finally, SF calculates the attention score, and combined with the spatial weight W_{sp}, we adjust the contribution of each spatial position. The concrete attention output is defined as Eq. (8–9).

$$F_{att} = Softmax\left(\frac{Q' \cdot K'^{T}}{\sqrt{d_k}}\right)V \tag{8}$$

$$F_{mod} = F_{att} \odot W_{sp} \tag{9}$$

$$O(SF) = (N)^2 \bullet 1 + (N)^2 \bullet C, O(SRA) = (N)^2 \bullet C + (N)^2 \bullet C \tag{10}$$

where *Softmax* normalization operations are performed by spatial location, F_{att} represents the output of attention, F_{mod} represents the weighted attention output. The above derivations show that SF achieves an effect similar to the full attention mechanism by first compressing the high-dimensional channels and then using adaptive spatial weight adjustment, while greatly reducing the computational overhead. Moreover, the computational complexity of SF can be expressed as Eq. (10). Compared to the previously lightweight and effective SRA, SF reduces the total computational cost of the attention operation by approximately half, strongly supporting the practical application and deployment of depth estimation tasks.

In order to prove that the SF designed can retain the key information even after dimensionality reduction, and can compensate the local spatial details that may be lost during dimensionality reduction by proper spatial weighting. We further define that for the i-th and j-th token in the input sequence, we have the high-dimensional representation $T_i, T_j \in \mathbb{R}^{d_k}$. In the standard attention mechanism, the similarity between the two is calculated as a normalized inner product, as shown in Eq. (11).

$$sim(i,j) = \frac{1}{\sqrt{d_k}} T_i \bullet T_j^T = \frac{1}{\sqrt{d_k}} \sum_{k=1}^{d_k} T_i(k) T_j(k) \tag{11}$$

where $T_i(k)$ and $T_j(k)$ represents the k-th component of the vector T_i and T_j. The operation of the compressed mapping function P can be written in the form of Eq. (12).

$$P(T_i) = \sum_{k=1}^{d_k} \alpha_k T_i(k), P\left(T_j\right) = \sum_{k=1}^{d_k} \beta_k T_j(k) \tag{12}$$

$$T_i' T_j' = \sum_{k=1}^{d_k} \sum_{k=1}^{d_k} \alpha_k \beta_k T_i(k) T_j(k) \tag{13}$$

where α_k and β_k are projected weights. After dimensionality reduction we get T_i' and T_j', which continue to do the inner product of Eq. (13). Then expand the above formula, to simplify the discussion, we assume that the projection design is "diagonalized", which can be approximately written as Eq. (14).

$$T_i' T_j' \approx \sum_{k=1}^{d_k} \alpha_k \beta_k T_i(k) T_j(k) \tag{14}$$

$$T_i' T_j' \approx \sum_{k=1}^{d_k} w_k T_i(k) T_j(k) \tag{15}$$

$$sim'(i,j) = \frac{1}{\sqrt{d_k}} T_i' T_j'^{T} = \frac{1}{\sqrt{d_k}} \sum_{k=1}^{d_k} w_k T_i(k) T_j(k) \tag{16}$$

We define the weighting factor on each channel $w_k \triangleq \alpha_k \beta_k$, and then we get Eq. (15). This is exactly similar to the traditional form of cross-correlation or convolution, where w_k is equivalent to weighting the response of each component. That is, we get a weighted cross-correlation operation like Eq. (16). In addition, in order to enhance the differentiation of spatial position information, we introduce the spatial adaptive weight mapping W_{sp}, and apply the generated weights to the similarity after dimensionality reduction in the spatial dimension, so as to obtain Eq. (17).

$$sim_{mod}(i,j) = \frac{W_{sp}}{\sqrt{d_k}} T_i' T_j'^{T} = \frac{W_{sp}}{\sqrt{d_k}} \sum_{k=1}^{d_k} w_k T_i(k) T_j(k) \tag{17}$$

where $sim_{mod}(i,j)$ represents the similarity result of introducing spatial adaptive weights, and W_{sp} plays the role of differentiated weighting for each spatial position.

3.3. Decoder branch: metric depth estimation

Metric depth estimation is a key component of the Decoder branch, which consists of the WBM. This module achieves fine-grained metric depth estimation with physical scales by cleverly combining wavelet transform convolution and an excellent bin design.

The design of the WBM is partially inspired by (Bhat et al. 2021, Bhat et al. 2022, Bhat et al. 2023). However, in contrast to the prevailing tendency in the field to focus on the internal structure of the bins themselves, such as the generation, splitting, and movement of bin centres, the WBM adopts a different approach. Instead of simply receiving feature information from the network bottleneck as input, which exaggerates the information aggregation capability of the MLP for multi-scale features and overlooks the rich low-frequency object shape features contained in different scale features, we add a wavelet transform-based convolution (WTconv) at the front end of the WBM, along with residual connections(Finder et al., 2024). As far as we know, this is the first time that WTConv has been applied to Bin design. Unlike standard convolution that focuses on high-frequency textures, WTConv decomposes features into multi-frequency components and hierarchically fuses them, enhancing the ability of model to recognize object shapes. This is crucial in port scenarios where the surfaces of equipment, such as worn containers and terminal cranes, lack texture. In neural networks, a convolution kernel consisting of multiple predefined or learnable wavelet filters $\{\psi_i\}_{i=1}^{M}$ is usually used to realize the wavelet transform convolution. For the input feature graph $X \in \mathbb{R}^{H\times W\times C}$, let M be the number of filters (corresponding to different frequency subbands), and for each filter ψ_i calculate the convolution operation as shown in Eq. (18).

$$Y_i = X^* \psi_i, i = 1, \cdots, M \tag{18}$$

where * represents a two-dimensional convolution operation, Y_i represents the characteristic response of the i-th frequency subband. In order to fuse the information of each frequency subband, we define a set of learnable fusion coefficients γ_i, and carry out weighted summation of each subband response to get Eq. (19).

$$X_{WT} = \sum_{i=1}^{M} \gamma_i Y_i \tag{19}$$

where X_{WT} is the multi-frequency feature map after fusion. For the untextured region, the image is mainly composed of low-frequency information, in which case $Y_{low} = X^* \psi_{low}$. That is, the response obtained by the low-pass filter is strong, while the response obtained by the high-frequency filter is weak, which can be automatically adjusted by γ_i, so that the low-frequency part plays a dominant role in the fusion, and finally the strong characteristic response to the untextured region is obtained, such as Eq. (20), which effectively enhances the shape information of the untextured region. This is very important to improve the reliability of AGV visual navigation system in port environment.

$$X'_{WT} = \gamma_{low} Y_{low} + \sum_{i \neq low} \gamma_i Y_i \tag{20}$$

In WBM, the fused feature X_{WT} will be used as the input of the subsequent Bin module to generate bin embeddings. This step is implemented by the Multi-Layer Perceptron (MLP), which maps X_{WT} to the logits predicted by bin, as shown in Eq. (21).

$$Z = MLP(X_{WT}) \tag{21}$$

$$P(k) = \frac{\exp(z_k)}{\sum_{j=1}^{B} \exp(z_j)}, k = 1, \cdots, B \tag{22}$$

$$D(i) = \sum_{k=1}^{B} P(k) \bullet c_k \tag{23}$$

where $Z \in \mathbb{R}^B$, B represents the total number of bins. Here, each component z_k represents the unnormalized probability of the model in that bin. Next, the logits are converted to a normalized probability distribution using the softmax function, as shown in Eq. (22). $P(k)$ represents the weight of the k-th bin under the current feature condition, which implicitly reflects the contribution of multi-frequency features to bin selection. Define the central value of each bin as c_k, these central values are derived from the uniform segmentation of the maximum and minimum depths in the bin chain in Fig. 3., and the final depth prediction $D(i)$ at each pixel position is given by the weighted average of the bin centers, as shown in Eq. (23). The above process shows how WTConv can fuse responses of different frequencies and use weighted bin embedding to integrate multi-frequency information into depth prediction, especially in untexturized regions where low frequency information plays a dominant role to achieve more robust depth estimation.

3.4. Decoder branch: relative depth estimation

The other branch in the Decoder stage is the relative depth estimation branch, which consists of the recovery model and the feature fusion module. It is used to recover the relative depth information between pixels in the image from the backbone.

The depth estimation task ultimately requires outputting the real-world distance information for each pixel in the image. Therefore, it belongs to the category of dense prediction tasks, and the final result must be a depth map rather than scattered image tokens. Consequently, it is logical to consider the necessity to re-aggregate the tokens that have been processed by the GSB into an image-like representation. To achieve this goal, we propose a Recovery Module following the approach of Dosovitsky et al (Dosovitskiy et al., 2020).

As shown in Fig. 4. First, we mix the previously added class-token into the other tokens and use linear projection to keep the dimension D unchanged, as shown in Eq.(24). The reason for not simply ignoring the class token is that the process from pixels to bin centers in the WBM can also be abstractly viewed as a classification task, and retaining the class token may be better for the overall task. Then, based on the position embedding that implicitly contains the initial pixel spatial location information, we concatenate the N tokens in an ordered manner to form an image-like feature map, as shown in Eq.(25). Finally, we use a convolutional layer to uniformly scale the feature map to a size of $(h/r) \times (h/r) \times \hat{D}$, where $\hat{D}$ is set to 256, as shown in Eq. (26).

$$Z'_l = Proj(concat(Z_l)), Z_l \in \mathbb{R}^{(N+1)\times D}, Z_{l'} \in \mathbb{R}^{N \times D} \tag{24}$$

$$Z'' = Link(Z'_l), Z'' \in \mathbb{R}^{\frac{h}{p} \times \frac{w}{p} \times D} \tag{25}$$

$$\hat{Z} = Resample(Z''), \ \hat{Z} \in \mathbb{R}^{\frac{h}{r} \times \frac{w}{r} \times \hat{D}} \tag{26}$$

where Z_l', Z'' and $\hat{Z}$ are all intermediate variables of image tokens. For the Feature Fusion Module, we simply employ a combination of residual convolutions and upsampling layers, with each level upsampling by a factor of 2. Finally, we design a task-specific head to generate the final prediction results.

3.5. Training strategies and loss function

As previously stated, the final prediction of our proposed MFMDepth is derived from a combination of relative depth and metric depth. Consequently, our training is carried out in two stages. Firstly, we remove the WBM and only train the backbone network for relative depth estimation. This training stage utilizes the Karlsruhe Institute of Technology and Toyota Institute of Technology 3D Object Detection Dataset (KITTI) dataset and the indoor depth dataset from New York University (NYU Depth v2) (Geiger et al. 2012, Silberman et al. 2012). During the training process in this stage, the backbone network learns the ability to predict depth at different scales. Subsequently, we simultaneously train our backbone network and WBM on depth data from various scenes in the port environment to accomplish the monocular metric depth estimation task for that particular scene. The construction of the port environment depth dataset will be elaborated in subsection 4.2.

We employ the structural similarity index measure loss (SSIM) for pixel-level supervised training. SSIM is commonly used to measure the structural similarity between the predicted depth map and the ground truth depth map. It is achieved by calculating the magnitude of differences in brightness, contrast, and structure between the two images. Its mathematical expression is given in Eq. (27) and Eq. (28).

$$L_{SSIM(y,\hat{y})} = -\log\left(\frac{1}{2|P|}\sum_{p \in P(y,\hat{y})}(1 + SSIM(p))\right) \tag{27}$$

$$SSIM(y,\hat{y}) = \frac{\left(2\mu_y\mu_{\hat{y}} + C_1\right)\left(2\sigma_{y\hat{y}} + C_2\right)}{\left(\mu_y^2 + \mu_{\hat{y}}^2 + C_1\right)\left(\sigma_y^2 + \sigma_{\hat{y}}^2 + C_2\right)} \tag{28}$$

The process of simplifying the weighted product of luminance,

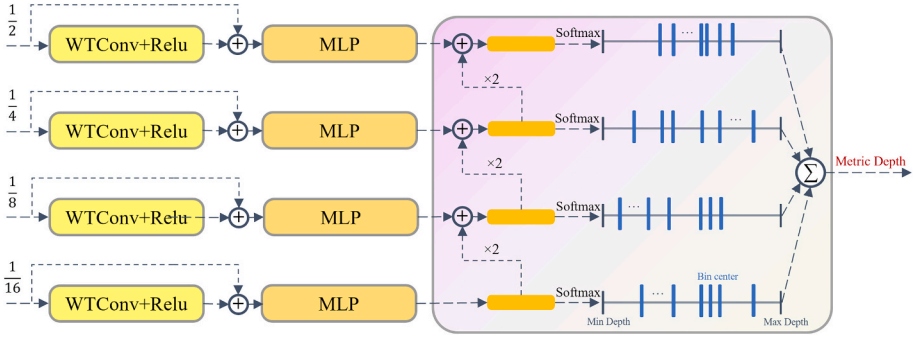

Fig. 3. The structure of WBM. The input are the image-like features of different resolutions at the bottlenecks. The weighted average of the centers of different bins is performed to obtain the final measurement depth result.

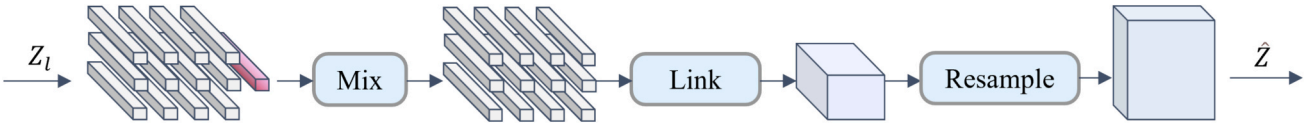

Fig. 4. Recovery module design structure.

contrast and structure to SSIM is simplified here. Where P represents the computed pixel level, with a value of 1 indicating per-pixel computation. $\hat{y}$ and y denote the predicted depth and ground truth. μ represents the mean, σ represents the variance, and $\sigma_{y\hat{y}}$ represents the covariance. C_1 and C_2 are stabilizing constants used to prevent division by zero. They are typically defined as $C_1 = (k_1 \cdot L)^2$ and $C_2 = (k_2 \cdot L)^2$ where k_1 is set to 0.01, k_2 is set to 0.03, and L represents the dynamic range of the pixel values.

4. Experiments

4.1. Evaluation metrics

To measure the performance of models in dense prediction tasks such as depth estimation, it is necessary to calculate the deviation between the depth predicted by the model for each pixel and the ground truth depth captured by the camera. We use the absolute relative error (AbsRel), root mean square error (RMSE), and $RMSE_{log}$ to quantify the error, and the threshold accuracy $\delta_n = \%$ of pixels to evaluate the performance of different models (Arampatzakis et al., 2023).

$$AbsRel = \frac{1}{N}\sum \frac{|y_i - \hat{y}_i|}{y_i} \tag{29}$$

$$RMSE = \sqrt{\frac{1}{N}\sum |y_i - \hat{y}_i|^2} \tag{30}$$

$$RMSE_{log} = \sqrt{\frac{1}{N}\sum |\log y_i - \log \hat{y}_i|^2} \tag{31}$$

$$\delta_n = \left[max\left(\frac{y_i}{\hat{y}_i}, \frac{\hat{y}_i}{y_i}\right) < thr^m \right] \%, thr = 1.25, m = 1, 2, 3 \tag{32}$$

where N is the total number of pixels, y_i is the ground truth, $\hat{y}_i$ is the predicted depth, and i is the index of pixel coordinates. The threshold accuracy calculates the percentage of pixels whose ratio between the predicted and true depth values is less than the threshold thr, with a value closer to 1 indicating better performance of the model.

4.2. Datasets

To evaluate the performance of our proposed framework for monocular depth estimation in port environments, we independently collected RGB-Depth image pairs (i.e., each frame of color images and corresponding depth images) from different scenes in port environments to construct an evaluation dataset. The camera model used in this study is the Orbbec Astra S, with a sampling resolution of 640 × 480 and a frame rate of 30 fps during the acquisition process. Camera calibration is an important prerequisite for visual measurement and positioning, and the accuracy of calibration parameters directly affects the precision of the acquisition system(Zhang, 2002). We employed the Zhang calibration method (Zhang 2021), using a planar checkerboard to calibrate the imaging parameters of the depth camera, and then utilized the obtained intrinsic and extrinsic parameters to correct the distortion generated during camera imaging. In the port environment, we collected a total of 2,431 RGB-Depth image pairs of common port scenes (as shown in Fig. 5.), including the placement of (single/multiple) cone barrel obstacles on port roads, autonomous vehicle driving in the port, and container placement in the yard. These image pairs were divided into training, validation, and test sets in a ratio of 6:2:2.

When the maximum depth of the acquisition scene exceeds the limit acquisition distance specified by the hardware specifications of the depth camera, the acquired depth data will be distorted. As a result, the acquired depth value is NAN (invalid value) or 0 value, and multiple holes will appear when the depth map of the scene is visualized. In order to obtain high quality depth data, we traverse and collect each depth map data for depth completion. Median filtering is first used to remove noise and isolated pixels, providing a cleaner data base for subsequent hole identification and completion. The depth map is then converted into a binary image, using numerical judgment to label invalid pixels as 1 (holes) and valid pixels as 0. Then, the image morphological dilation operation is used to expand the edge of the hole to ensure the smooth boundary in subsequent interpolation, while the corrosion operation is used to remove small isolated invalid points to avoid misjudgment, and the depth value of the adjacent effective pixels is used to fill the hole. Finally, the depth data after repair are checked and annotated manually

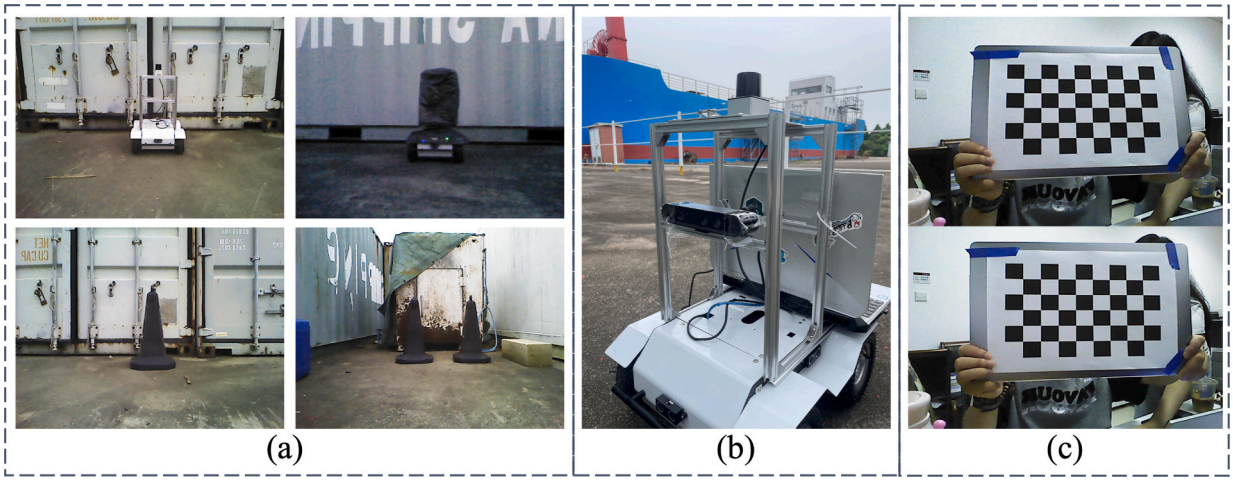

Fig. 5. Datasets compose structure. (a) presents some sample scenes from the dataset. (b) illustrates the hardware structure of the acquisition device. (c) shows the camera calibration checkerboard.

to correct the unreasonable depth value.

4.3. Parameters Setting

In the initial stage of the framework, where a single RGB image is reshaped into image patches, we set P to 16, the projection dimension D to 768, and the number of attention heads to 12 to obtain sufficient global relationship modeling performance (Ranftl et al. 2021). The AdamW optimizer, which separates weight decay from gradient updates, enables the network to acquire stronger generalization ability and improves training efficiency (Loshchilov & Hutter, 2017). We easily integrate it into our framework and set the initial learning rate within the range of $\{5e^{-5}, 1e^{-4}, 5e^{-4}, 1e^{-3}\}$. In the recovery module, we set the scale r of the resampling step to 2 to 16, with a scale factor of 2, to restore the image feature representation at different scales. For a fair comparison, all models used in the comparative experiments are trained on the same computing cluster, which contains 2 NVIDIA RTX A4000 and 1 NVIDIA RTX A5000. We set the maximum number of training epochs to 250 and employ an early stopping strategy, where training is terminated early when the model performance on the validation set no longer improves, to prevent model overfitting.

4.4. Experimental results and analysis

To verify the superiority of our proposed framework, we selected state-of-the-art monocular metric depth estimation methods in recent years for comparative experiments. Including Local Bins, AdaBins, and ZoeDepth, which have excellent bins module designs. DS-SIDE, which shares the same two-stage network design. MiDas, a lightweight network from Intel Labs focusing on real-time depth estimation tasks. DepthPro is a high-speed and accurate depth estimation method for outdoor scenarios proposed by Apple. The latest DAC, suitable for different industrial scenarios (Ren et al. 2019, Bhat et al. 2021, Bhat et al. 2022, Ranftl et al., 2020, Bhat et al. 2023, Bochkovskii et al. 2024, Guo et al. 2025). To ensure fairness in the experiments, the parameter settings for all comparative models are consistent with the experimental parameters in their original papers.

We simulate two common scenarios in the port environment: an AGV encountering obstacles while driving and two AGVs carrying containers in opposite directions. In the first scenario, the AGV approaches a traffic cone during its journey, perceives the obstacle, and then moves away from it. This scenario has a relatively small overall depth range, similar to an indoor scene. The depth camera is mounted on the front of the AGV, and the predicted depth by the model for the traffic cone should gradually decrease and then increase. The second scenario is set in a port yard environment where two AGVs drive towards each other, with the distance between them first decreasing and then increasing. This scenario maintains a large depth range. Based on the segmentation masks and depth maps obtained from the predictions of different heads in the model, the ground truth depth of specific objects in the scene can be derived from the weighted average of the metric depth prediction values within the mask range. To ensure fairness, all depth estimation models use the same mask.

To fully evaluate the performance of the depth estimation framework, we conducted an extensive comparative analysis in two different scenarios. Fig. 6(a), (b), (d) and (e) show the timing prediction accuracy of the actual distance values of the left and right traffic cones in the process of AGV movement. Our analysis reveals three key observations: (1) The proposed model consistently maintains a sub-decimeter prediction error (0.03–0.1 m) throughout the motion sequence and exhibits superior temporal stability compared to the baseline. (2) While ZoeDepth has the second-best performance, its error variance is from 12 % to 15 % higher for sudden changes in speed, likely due to its reliance on the distribution of pre-trained data. (3) Both the AdaBins and the LocalBins show significant error accumulation (e.g., up to 0.21 m offset at frame 380 of Fig. 6(e)), suggesting that their Bin-based designs have difficulty with surface features that are lacking in texture in a port environment. However, DS-SIDE, which is based on the traditional CNN network, lacks the global modeling ability of low-frequency features, resulting in inaccurate depth prediction.The complementary perspectives in Fig. 6(c) and (f) highlight the advantages of our model in full-scene depth perception. Particularly in Scenario 2 containing complex container arrangements, our framework achieves a depth prediction RMSE approximately 26.3 % lower than MiDas. This improvement is the result of our novel hybrid architecture, which collaboratively combines WT Bins modules for low-frequency feature preservation and high-frequency detail extraction, a feature conspicuously lacking in traditional DS-SIDE CNN designs.

Fig. 7. provides qualitative validation through continuous depth map predictions. The unique advantages of our approach can be seen from the visual comparison. First, our model maintains spatial consistency in the transition region, especially in the smooth depth gradient between stacked containers (e.g., the depth prediction of the box behind the cone and the gap between containers on both sides in the first scenario). In addition, compared to other models, our model maintains fine-grained depth predictions at the edge of obstacles, such as the canvas edge on

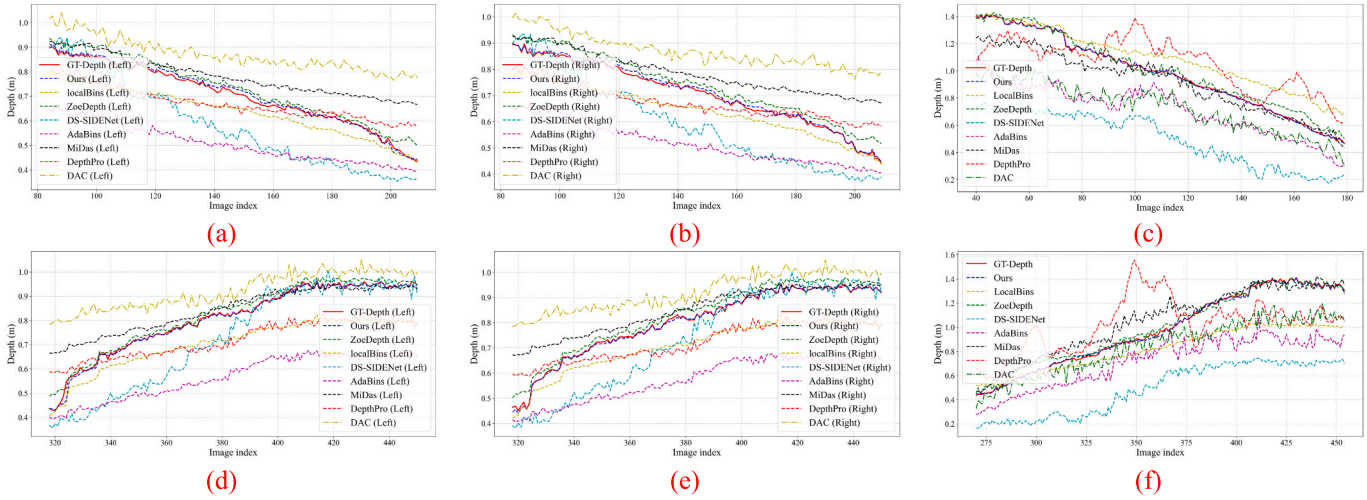

Fig. 6. Prediction of the real distance to the traffic cones. (a), (b), (d), and (e) depict the depth prediction for the left and right traffic cones. (c) and (f) illustrate the depth prediction results for the moving AGV. The red solid line represents the ground truth collected by the depth camera, while the other dashed lines indicate the prediction results of different models.

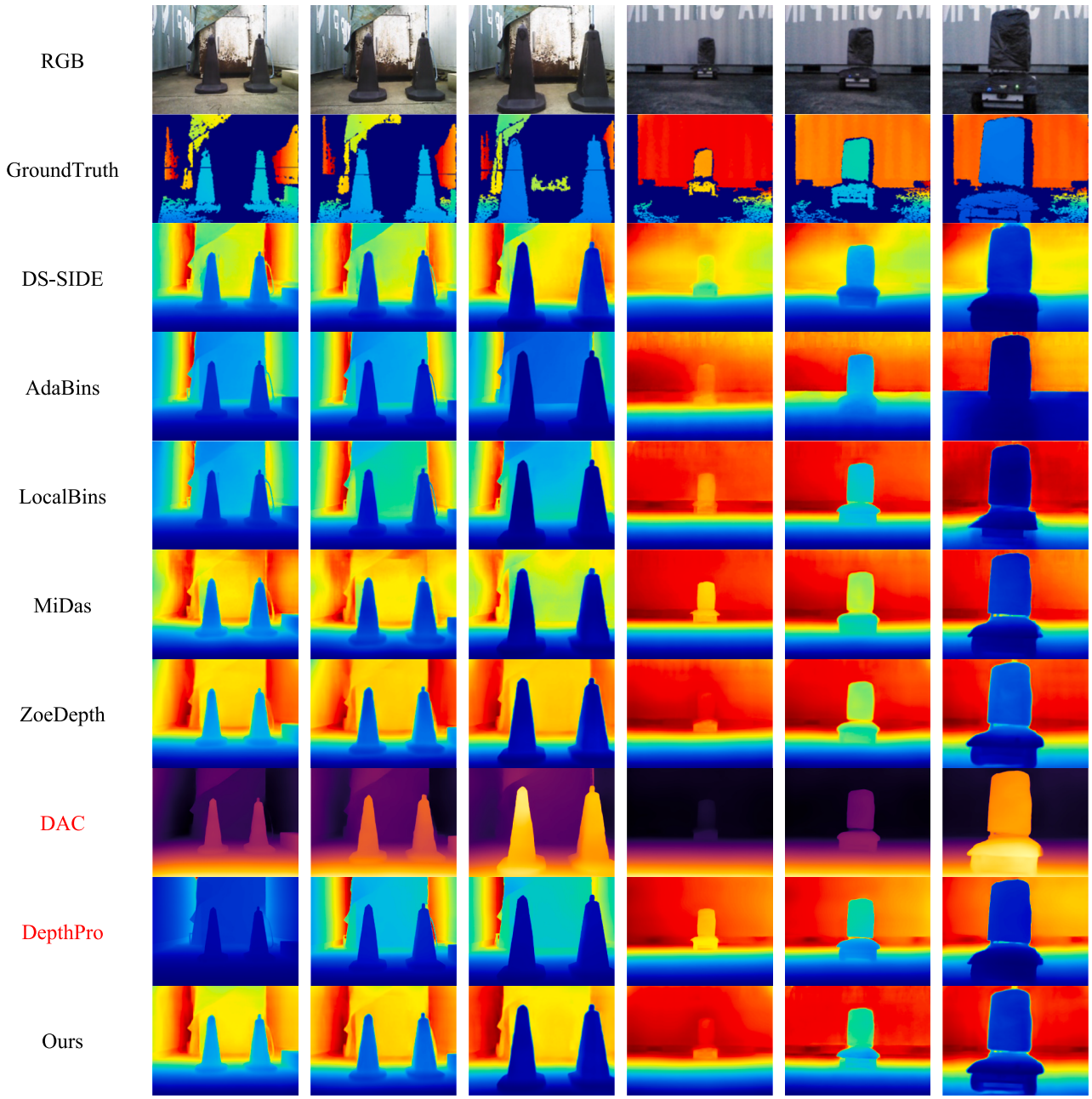

Fig. 7. Visualization of the metric depth estimation by models in the port scenario. Different color tones correspond to the magnitude of the estimated metric depth values.

the box behind the cone and the pipe protruded to the right in the first scenario, and the cavity below the container behind the AGV in the second scenario.

Table 2 and Table 3 show the comparison results of quantitative metrics between our model and other models in the two scenarios, respectively. It can be observed that our proposed model significantly

Table 2
Performance comparison on traffic cone scene. Measurements are made for the depth range from 0.2 m to 10 m. The best results are shown in bold.(↑ indicates that bigger is better, ↓ indicates that the smaller the better).

Model Name	AbsRel ↓	RMSE ↓	$RMSE_{log}$ ↓	$\sigma < 1.25$↑	$\sigma < 1.25^2$↑	$\sigma < 1.25^3$↑
DS-SIDE	0.131	0.555	0.053	0.841	0.953	0.988
AdaBins	0.110	0.392	0.047	0.885	0.978	0.994
LocalBins	0.099	0.351	0.043	0.907	0.986	0.998
MiDas	0.082	0.294	0.035	0.946	0.994	0.997
ZoeDepth	0.075	0.270	0.032	0.955	0.995	0.999
DepthPro	0.063	0.251	0.028	0.975	0.994	0.998
DAC	0.067	0.260	0.030	0.980	0.997	0.999
Ours	**0.059**	**0.206**	**0.024**	**0.984**	**0.998**	**1.000**

Table 3
Performance Comparison on AGV Driving Scenario Data. Measurements are made for the depth range from 0.2 m to 60 m. The best results are shown in bold.

Model Name	AbsRel ↓	RMSE ↓	$RMSE_{log}$ ↓	$\sigma < 1.25$↑	$\sigma < 1.25^2$↑	$\sigma < 1.25^3$↑
DS-SIDE	0.114	4.935	0.206	0.861	0.949	0.976
AdaBins	0.098	3.933	0.173	0.890	0.964	0.985
LocalBins	0.072	2.727	0.120	0.932	0.984	0.994
MiDas	0.062	2.573	0.092	0.959	0.995	0.998
ZoeDepth	0.048	2.045	0.072	0.976	0.997	0.999
DepthPro	0.058	2.770	0.092	0.964	0.993	0.998
DAC	0.051	2.403	0.080	0.977	0.996	0.999
Ours	**0.046**	**1.896**	**0.069**	**0.982**	**0.999**	**1.000**

outperforms the SOTA models in all metrics, and the substantial performance gap with the previous state-of-the-art models emphasizes the effectiveness of our model architecture design. Specifically, compared to the previous DS-SIDE, our model achieves a performance improvement of nearly twofold. Compared to the current SOTA Zoedepth, our model improves the RMSE by 23.7 % and the AbsRel by 21.3 %. Moreover, in scenario two, which has a larger depth variation range, our model also achieves an improvement of approximately 7.2 % in RMSE compared to Zoedepth. In terms of the threshold accuracy metric, only our model achieves a perfect accuracy of 1.000 when $\delta_n < 1.25^3$.

This comprehensive evaluation confirms that our architecture successfully addresses three persistent challenges of depth estimation methods in automated pier scenarios: time consistency maintenance, texture-agnostic feature learning and high-precision edge recovery. The significant performance gap (at least 21.3 % improvement) over existing methods particularly highlights the effectiveness of our frequency-aware and relative/metric deep fusion learning paradigms in AGV production environments.

5. Ablation study

5.1. Effectiveness of squeeze former module

Table 4 validates the effectiveness of using the Squeeze Former as the token mixer. We conducted experiments on cases where SF is applied or not applied in the Global Squeeze Block, with traditional attention mechanisms replacing SF when not used. The results show that using SF throughout the GSB achieves an improvement of approximately 14.5 % in RMSE compared to using traditional attention mechanisms entirely in the GSB. As SF is gradually added to each layer in the GSB, the AbsRel metric also decreases, indicating that using SF as the token mixer can effectively capture global context from global features, thus continuously improving model performance.

Fig. 8. visualizes the comparison of attention distributions between using SF and traditional attention mechanisms in the GSB_3 stage. The results demonstrate that our proposed SF exhibits competitive attention distribution performance while reducing computational consumption. In contrast, the attention distribution of traditional attention mechanisms becomes scattered in both image two and four, whereas our SF still performs well. Table 5 shows the inference time of different models. The "parameters" parameter is used to calculate the total number of all trainable weights in the model, and the calculation method follows the principle of layer by layer accumulation. The calculation methods of parameters of different network layers in the model are given below.

Table 4
Ablation study for applying our proposed SF to different layer.

GSB_2	GSB_3	GSB_4	NYU	
			RMSE	AbsRel
×	×	×	0.316	0.105
✓	×	×	0.301	0.089
✓	✓	✓	0.281	0.081
✓	✓	✓	**0.270**	**0.072**

$$P_{BN} = 2c \tag{33}$$

$$P_{FC} = (d_{in} \times d_{out}) + d_{bias} \tag{34}$$

$$P_{Conv} = (k_w \times k_h \times C_{in} \times C_{out}) + C_{out} \tag{35}$$

$$P_{Trans} = l\left(12h^2 + 11h + 4c\right) \tag{36}$$

where P_{FC}, P_{Conv}, P_{Trans} and P_{BN} represent parameter number statistics of fully connected layer, convolution layer, transformer module and normalized layer, respectively; d_{in} is the input dimension, d_{out} is the output dimension, and d_{bias} is the bias dimension; k_w and k_h are the dimensions of the convolution kernel, and C_{in} and C_{out} are the number of input and output channels of the convolution layer; l represents the number of layers of the transformer module, h is the dimension of the attention matrix, and c is the number of channels in the normalized layer. Timings were conducted on an Intel Core i7-10700 K CPU @ 3.80 GHz with 16 physical cores and an NVIDIA RTX A4000. We use square images with a width of 480 pixels and report the average over 250 runs. As demonstrated in Table IV, while our model incorporates a substantially greater number of parameters in comparison to alternative model architectures, leveraging the high parallelism intrinsic to SF and the mitigation of computational overhead in the Query and Key matrix channels, it attains a latecy that is analogous to that of MiDas, which employs a fully CNN architecture.

5.2. Effectiveness of WT bins module

To intuitively understand the gain of receptive field by integrating convolution based on wavelet transform (WTConv) in WBM module and overcome the deficiency of traditional Conv2d only responding to high-frequency information, we conducted experiments on the response of high and low-frequency signals using a single layer of WTConv and Conv2d at the same position in the model. Pulse signals and Gaussian signals were generated at the center position of the image, respectively. After the signals completely passed through the WTConv layer and Conv2d layer separately, the non-zero response area was calculated using a threshold (set to $1e^{-6}$ here) to obtain the receptive field size of the current layer.

As shown in Fig. 9 (b), (c). The receptive field size of WTConv is 64, covering the entire feature map with responses, whereas the receptive field of Conv2d is only 6, primarily focused in the central region of the image. This difference, nearly tenfold for high-frequency inputs, highlights a significant advantage in receptive field size. When subjected to low-frequency Gaussian signals, the receptive field size of standard convolution remains largely unchanged, while WTConv captures features across varying frequencies, demonstrating both positive and negative responses in Fig. 9(f). In port environments, automated equipment such as idle containers and quay cranes often exhibit blurred surface textures due to wear and tear. The strong ability of WTConv to capture low-frequency information allows for the accurate extraction of shape information, effectively compensating for the adverse effects caused by texture loss. Fig. 10 illustrates a depth prediction comparison

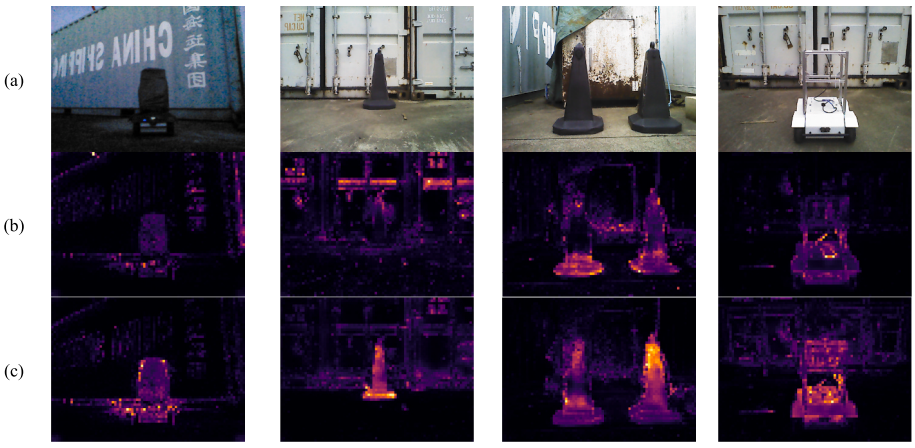

Fig. 8. Comparison of attention distributions. (a) shows the original images from the port depth dataset, (b) represents the attention distribution of traditional attention mechanisms, and (c) illustrates the attention distribution of SF.

Table 5
Inference Speed Comparison.

Model Name	MiDas	LocalBins	ZoeDepth	Ours
Parameters(m)	**109**	123	156	323
Time(ms)	30	**16**	39	33

between our model and Midas and ZoeDepth in textureless background regions and along rope edge details. The results show that our model provides more accurate depth estimation of object shapes and edges, demonstrating a clear advantage in fine-grained depth estimation performance.

6. Conclusion

This study proposes MFMDepth, a novel monocular metric depth estimation framework designed to address the critical challenges of visual navigation for port AGVs. By integrating the Global Squeeze Block (GSB) and Wavelet-based Bins Module (WBM) into a unified Metaformer architecture, our framework achieves three key advancements. First, the GSB module resolves the computational inefficiency of conventional Transformer-based encoders through dual-channel compression and adaptive spatial weighting. This breakthrough enables real-time processing (33 ms/frame) on embedded AGV systems, fulfilling the stringent latency requirement for collision avoidance in dynamic port environments. Second, the WBM module introduces a wavelet-driven metric depth estimation paradigm to overcome texture ambiguity in structurally homogeneous scenes. By decomposing depth features into multi-frequency components via wavelet transforms, WBM achieves a 21.3 % improvement in boundary estimation accuracy (measured by MAE) for stacked containers compared to ZoeDepth's Bins designs. Finally, the dual-stream fusion architecture bridges the gap between relative and metric depth estimation. Experimental results on real port

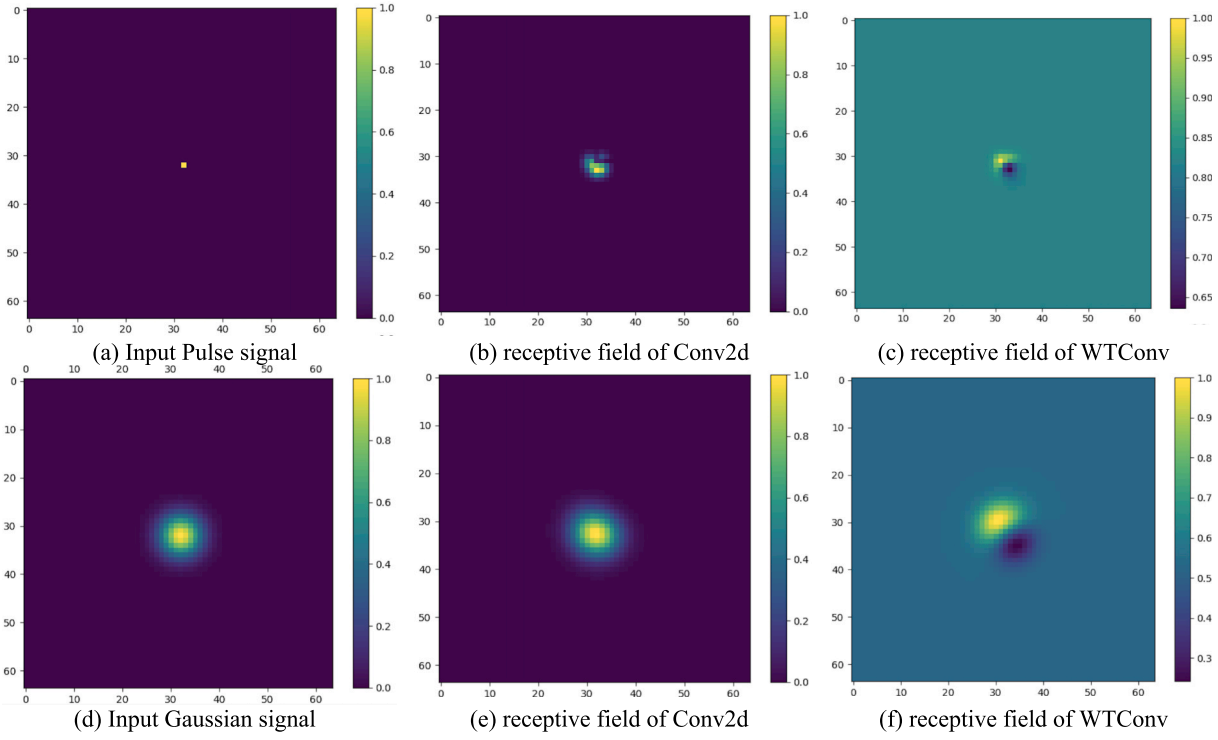

Fig. 9. Visualization of the receptive field sizes of single-layer convolution input pulses and Gaussian signals.

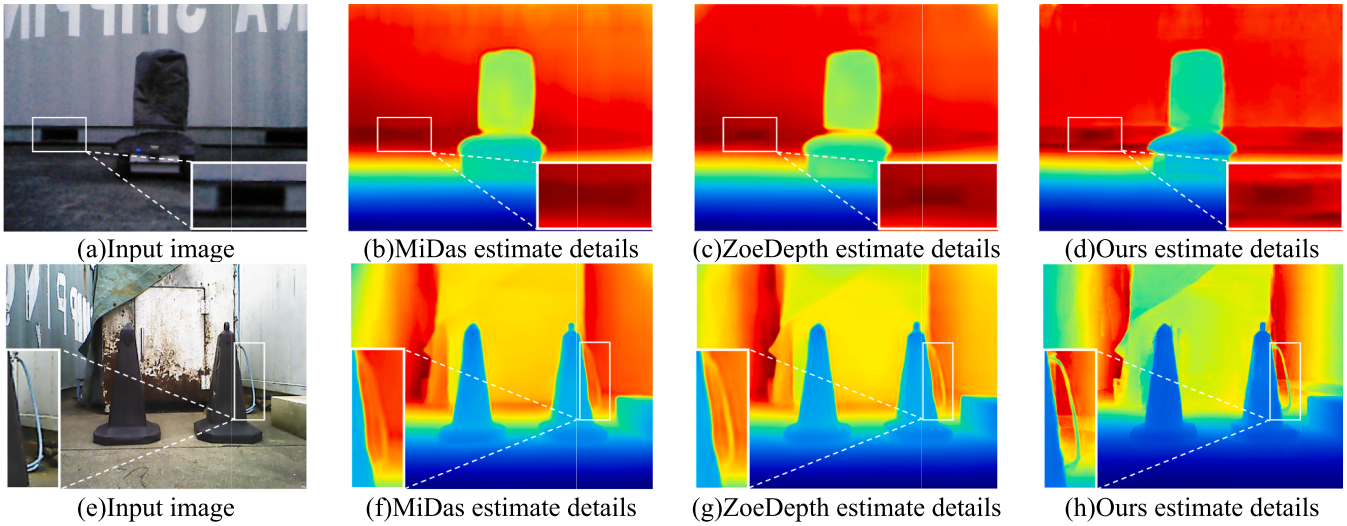

Fig. 10. Comparison of the effect of depth estimation on the edge details of objects.

datasets show new SOTA performance, with AbsRel/RMSE/RMSElog reaching 0.059/0.206/0.024, outperforming existing SOTA methods by from 21.3 % to 25 %. Beyond technical contribution, this work also provides a practical solution to the annual high maintenance costs of magnetic nail systems.

In future work, we plan to explore the zero-shot depth estimation capabilities of MFMDepth and integrate our model with large language models to further enhance its adaptability and generalization across diverse port environments without additional retraining. Specifically, the integration with large language models (LLMs) will be implemented via a cross-modal attention framework, where linguistic embeddings from port operation manuals and equipment descriptions (processed by LLMs) will interact with visual features through attention gates. This architecture draws inspiration from recent success in SpatialVLM and LLM-MDE for joint vision-language representation learning, allowing depth estimation networks to leverage semantic context (e.g., container crane in foreground or ship berthing area) to resolve geometric ambiguities (Chen et al. 2024, Xia & Wu, 2024). Preliminary experiments using BERT-encoded cargo handling instructions have shown 12 % improvement in relative depth ordering accuracy on unseen port layouts. Another direction is to expand the model's depth estimation ability under challenging conditions (e.g., low-light and adverse weather scenarios) by integrating domain adaptation and robustness enhancement techniques and considering the sensitivity influence of each parameter (e.g., the correlation between different receptive field sizes and edge prediction accuracy, etc.).

CRediT authorship contribution statement

Xinqiang Chen: Writing – review & editing, Conceptualization. **Fei Ma:** Writing – original draft, Visualization, Methodology. **Yuzheng Wu:** Data curation. **Bing Han:** Funding acquisition, Formal analysis. **Lijuan Luo:** Project administration. **Salvatore Antonio Biancardo:** Software.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was jointly supported by National Natural Science Foundation of China (52331012, 52472347, 52071200). Open Fund of Chongqing Key Laboratory of Green Logistics Intelligent Technology (Chongqing Jiaotong University) (No.KLGLIT2024ZD001). Shanghai Committee of Science and Technology, China (23010502000).

Data availability

The data that has been used is confidential.

References

Aizat, M., Azmin, A., & Rahiman, W. (2023). A survey on navigation approaches for automated guided vehicle robots in dynamic surrounding. *IEEE Access, 11*, 33934–33955.

Arampatzakis, V., Pavlidis, G., Mitianoudis, N., & Papamarkos, N. (2023). Monocular depth estimation: A thorough review. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 46*(4), 2396–2414.

Bhat, S. F., Alhashim, I., & Wonka, P. (2021). Adabins: Depth estimation using adaptive bins. In *In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 4009–4018).

Bhat, S. F., Alhashim, I., & Wonka, P. (2022). In *Localbins: Improving depth estimation by learning local distributions* (pp. 480–496). Cham: Springer Nature Switzerland.

Bhat, S. F., Birkl, R., Wofk, D., Wonka, P., & Müller, M. (2023). Zoedepth: Zero-shot transfer by combining relative and metric depth. arXiv preprint arXiv:2302.12288.

Bochkovskii, A., Delaunoy, A., Germain, H., Santos, M., Zhou, Y., Richter, S. R., & Koltun, V. (2024). Depth pro: Sharp monocular metric depth in less than a second. arXiv preprint arXiv:2410.02073.

Chen, B., Xu, Z., Kirmani, S., Ichter, B., Sadigh, D., Guibas, L., & Xia, F. (2024). Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 14455–14465).

Chen, D., Huang, C., Fan, T., Lau, H. C., & Yan, X. (2025). Predictive modeling for vessel traffic flow: A comprehensive survey from statistics to AI. *Transportation Safety and Environment, tdaf022*.

Chen, S., Cheng, K., Yang, J., Zang, X., Luo, Q., & Li, J. (2023). Driving behavior risk measurement and cluster analysis driven by vehicle trajectory data. *Applied Sciences, 13*(9), 5675.

Chen, S., Piao, L., Zang, X., Luo, Q., Li, J., Yang, J., & Rong, J. (2023). Analyzing differences of highway lane-changing behavior using vehicle trajectory data. *Physica A: Statistical Mechanics and its Applications, 624*, Article 128980.

Chen, X., Hu, R., Luo, K., Wu, H., Biancardo, S. A., Zheng, Y., & Xian, J. (2025). Intelligent ship route planning via an A* search model enhanced double-deep Q-network. *Ocean Engineering, 327*, Article 120956.

Chen, X., Wu, H., Han, B., Liu, W., Montewka, J., & Liu, R. W. (2023). Orientation-aware ship detection via a rotation feature decoupling supported deep learning approach. *Engineering Applications of Artificial Intelligence, 125*, Article 106686.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

Finder, S. E., Amoyal, R., Treister, E., & Freifeld, O. (2024). In *Wavelet convolutions for large receptive fields* (pp. 363–380). Cham: Springer Nature Switzerland.

Geiger, A., Lenz, P., & Urtasun, R. (2012). In *Are we ready for autonomous driving? the kitti vision benchmark suite* (pp. 3354–3361). IEEE.

Guizilini, V., Ambrus, R., Pillai, S., Raventos, A., & Gaidon, A. (2020). 3d packing for self-supervised monocular depth estimation. In *In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 2485–2494).

Guo, Y., Garg, S., Miangoleh, S. M. H., Huang, X., & Ren, L. (2025). Depth Any Camera: Zero-Shot Metric Depth Estimation from Any Camera. arXiv preprint arXiv: 2501.02464.

Hu, J., Fan, C., Zhou, L., Gao, Q., Liu, H., & Lam, T. L. (2023). Lifelong-MonoDepth: Lifelong Learning for Multidomain Monocular Metric Depth Estimation. *IEEE Transactions on Neural Networks and Learning Systems*.

Hu, M., Yin, W., Zhang, C., Cai, Z., Long, X., Chen, H., & Shen, S. (2024). Metric3d v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Huang, P., Xiong, Y., Tang, S., Wang, S., Zeng, Q., & Lee, J. J. (2025). Driver injury severity analysis of work zone crashes: A Bayesian hierarchical generalized ordered probit approach. *Transportation Safety and Environment*, Article tdaf001.

Huang, W., Deng, Y., Hui, S., Wu, Y., Zhou, S., & Wang, J. (2024). Sparse self-attention transformer for image inpainting. *Pattern Recognition, 145*, Article 109897.

Liu, F., Huang, M., Ge, H., Tao, D., & Gao, R. (2023). Unsupervised monocular depth estimation for monocular visual SLAM systems. *IEEE Transactions on Instrumentation and Measurement, 73*, 1–13.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., & Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *In Proceedings of the IEEE/CVF international conference on computer vision* (pp. 10012–10022).

Loshchilov, I., & Hutter, F. (2017). Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101.

Luo, X., Ai, Z., Liang, Q., Xie, Y., Shi, Z., Fan, J., & Qu, Y. (2024). *EdgeFormer: Edge-aware Efficient Transformer for image Super-resolution*. IEEE Transactions on Instrumentation and Measurement.

Mertan, A., Duff, D. J., & Unal, G. (2022). Single image depth estimation: An overview. *Digital Signal Processing, 123*, Article 103441.

Mohammed, A. A., Geng, X., Wang, J., & Ali, Z. (2024). Driver distraction detection using semi-supervised lightweight vision transformer. *Engineering Applications of Artificial Intelligence, 129*, Article 107618.

Poggi, M., Tosi, F., Batsos, K., Mordohai, P., & Mattoccia, S. (2021). On the synergies between machine learning and binocular stereo for depth estimation from images: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 44*(9), 5314–5334.

Ranftl, R., Bochkovskiy, A., & Koltun, V. (2021). Vision transformers for dense prediction. In *In Proceedings of the IEEE/CVF international conference on computer vision* (pp. 12179–12188).

Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., & Koltun, V. (2020). Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 44*(3), 1623–1637.

Raptakis, M., & Pantazis, Y. (2024). Fourier Attention: The Attention Mechanism as a Frequency Analyzer. In Proc. IberSPEECH 2024 (pp. 116-120).

Raza, Z., Woxenius, J., Vural, C. A., & Lind, M. (2023). Digital transformation of maritime logistics: Exploring trends in the liner shipping segment. *Computers in Industry, 145*, Article 103811.

Reis, W. P. N. D., Couto, G. E., & Junior, O. M. (2023). Automated guided vehicles position control: A systematic literature review. *Journal of Intelligent Manufacturing, 34*(4), 1483–1545.

Ren, H., El-Khamy, M., & Lee, J. (2019). Deep robust single image depth estimation neural network using scene understanding. *CVPR Workshops, 2*, p. 2).

Silberman, N., Hoiem, D., Kohli, P., & Fergus, R. (2012). Indoor segmentation and support inference from rgbd images. In Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part V 12 (pp. 746-760). Springer Berlin Heidelberg.

Sun, Q., Tang, Y., Zhang, C., Zhao, C., Qian, F., & Kurths, J. (2021). Unsupervised estimation of monocular depth and VO in dynamic environments via hybrid masks. *IEEE Transactions on Neural Networks and Learning Systems, 33*(5), 2023–2033.

Thwal, C. M., Nguyen, M. N., Tun, Y. L., Kim, S. T., Thai, M. T., & Hong, C. S. (2024). OnDev-LCT: On-device lightweight convolutional transformers towards federated learning. *Neural Networks, 170*, 635–649.

Venkataramanan, S., Ghodrati, A., Asano, Y. M., Porikli, F., & Habibian, A. (2023). Skip-attention: Improving vision transformers by paying less attention. arXiv preprint arXiv:2301.02240.

Wang, W., Xie, E., Li, X., Fan, D. P., Song, K., Liang, D., & Shao, L. (2021). Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *In Proceedings of the IEEE/CVF international conference on computer vision* (pp. 568–578).

Wu, X., Sun, C., Zou, T., Li, L., Wang, L., & Liu, H. (2020). SVM-based image partitioning for vision recognition of AGV guide paths under complex illumination conditions. robotics and computer-integrated manufacturing, 61, 101856.

Xia, Z., & Wu, T. (2024). Large Language Models Can Understanding Depth from Monocular Images. arXiv preprint arXiv:2409.01133.

Xiaogang, R., Wenjing, Y., Jing, H., Peiyuan, G., & Wei, G. (2020). In *Monocular depth estimation based on deep learning: A survey* (pp. 2436–2440). IEEE.

Yar, H., Khan, Z. A., Hussain, T., & Baik, S. W. (2024). A modified vision transformer architecture with scratch learning capabilities for effective fire detection. *Expert Systems with Applications, 252*, Article 123935.

Yin, W., Zhang, C., Chen, H., Cai, Z., Yu, G., Wang, K., & Shen, C. (2023). Metric3d: Towards zero-shot metric 3d prediction from a single image. In *In Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 9043–9053).

Yu, K., Li, H., Xing, L., Wen, T., Fu, D., Yang, Y., & Bai, H. (2023). Scene-aware refinement network for unsupervised monocular depth estimation in ultra-low altitude oblique photography of UAV. *ISPRS Journal of Photogrammetry and Remote Sensing, 205*, 284–300.

Yu, W., Luo, M., Zhou, P., Si, C., Zhou, Y., Wang, X., & Yan, S. (2022). Metaformer is actually what you need for vision. In *In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10819–10829).

Zhang, S., Zhang, J., & Tao, D. (2022). In *Towards scale-aware, robust, and generalizable unsupervised monocular depth estimation by integrating IMU motion dynamics* (pp. 143–160). Cham: Springer Nature Switzerland.

Zhang, Y., Liu, C., Liu, M., Liu, T., Lin, H., Huang, C. B., & Ning, L. (2024). Attention is all you need: Utilizing attention in AI-enabled drug discovery. *Briefings in Bioinformatics, 25*(1), Article bbad467.

Zhang, Z. (2021). Camera calibration. In *Computer vision: a reference guide* (pp. 130–131). Cham: Springer International Publishing.

Zhang, Z. (2002). A flexible new technique for camera calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 22*(11), 1330–1334.

Zhao, C., Sun, Q., Zhang, C., Tang, Y., & Qian, F. (2020). Monocular depth estimation based on deep learning: An overview. *Science China Technological Sciences, 63*(9), 1612–1627.

Zhu, Y., Ren, R., Dong, W., Li, X., & Shi, G. (2024). TSUDepth: Exploring temporal symmetry-based uncertainty for unsupervised monocular depth estimation. *Neurocomputing, 600*, Article 128165.