RIDERS: Radar-Infrared Depth Estimation for Robust Sensing

Han Li, Yukai Ma, Yuehao Huang, Yaqing Gu, Weihua Xu, Yong Liu, and Xingxing Zuo

***Abstract*— Dense depth recovery is crucial in autonomous driving, serving as a foundational element for obstacle avoidance, 3D object detection, and local path planning. Adverse weather conditions, including haze, dust, rain, snow, and darkness, introduce significant challenges to accurate dense depth estimation, thereby posing substantial safety risks in autonomous driving. These challenges are particularly pronounced for traditional depth estimation methods that rely on short electromagnetic wave sensors, such as visible spectrum cameras and near-infrared LiDAR, due to their susceptibility to diffraction noise and occlusion in such environments. To fundamentally overcome this issue, we present a novel approach for robust metric depth estimation by fusing a millimeter-wave radar and a monocular infrared thermal camera, which are capable of penetrating atmospheric particles and unaffected by lighting conditions. Our proposed Radar-Infrared fusion method achieves highly accurate and finely detailed dense depth estimation through three stages, including monocular depth prediction with global scale alignment, quasi-dense radar augmentation by learning radar-pixels correspondences, and local scale refinement of dense depth using a scale map learner. Our method achieves exceptional visual quality and accurate metric estimation by addressing the challenges of ambiguity and misalignment that arise from directly fusing multi-modal long-wave features. We evaluate the performance of our approach on the NTU4DRadLM dataset and our self-collected challenging ZJU-Multispectrum dataset. Especially noteworthy is the unprecedented robustness demonstrated by our proposed method in smoky scenarios. Our code will be released at https://github.com/MMOCKING/RIDERS.**

***Index Terms*— Depth estimation, radar perception, infrared camera, multi-sensor fusion.**

I. Introduction

PERCEPTION plays a critical role in autonomous driving and robotics, with depth estimation serving as the preliminary for dense reconstruction, obstacle avoidance, and 3D detection. Vehicles equipped with advanced driver assistance systems commonly utilize LiDAR, which operates in the near-infrared spectrum, and RGB cameras that capture the visible spectrum. This sensor fusion provides a comprehensive perception of complex 3D environments, combining accurate geometric range data with detailed visual information. Such integration is regarded as a dependable approach in various driving contexts.

However, due to the inherent limitations of LiDAR and RGB cameras, this combination can fail concurrently in adverse weather conditions like haze, dust, smoke, fog, rain, or snow. The visible light wavelength ranges from 390 nm to 780 nm, while mainstream LiDAR emits laser wavelengths of 905 nm and 1550 nm [1]. In the presence of micrometer-sized atmospheric particles, serious diffraction occurs, severely interfering with the imaging or echo reception. Existing solutions that enhance perceptual robustness at the software level [2], [3], [4] often face challenges in overcoming the inherent limitations of sensors. Methods that fit well on datasets or are effective against mild disturbances struggle to maintain performance in complex and dynamic real-world driving scenarios. Even the best denoising methods may fail in scenes with complete occlusion. Therefore, there is a demand for a robust depth estimation method that can fundamentally handle interferences such as smoke, rain, and snow. Addressing this challenge requires a perspective shift toward robust sensor solutions.

In the field of autonomous driving, mmWave radars and infrared thermal cameras, known for their high resilience in adverse weather conditions, are increasingly recognized. Automotive imaging radars, primarily operating at 77 GHz within the W-band [5], feature a wavelength of approximately 3.9 mm. On the other hand, thermal cameras generally operate within a wavelength range of 7-14 μm. Due to their minimal susceptibility to atmospheric particulates and variable lighting conditions, these sensors offer a viable alternative to the conventional LiDAR and RGB camera ensemble, especially in challenging environments such as nighttime, smoke-filled, or foggy conditions.

Current methodologies integrating radar and infrared camera technology predominantly focus on target detection and pose estimation tasks [6], [7], [8]. However, the potential for their fusion in achieving 3D dense depth estimation remains largely untapped. Similar to the limitations of monocular solutions with RGB cameras, depth estimation using a single infrared camera struggles to accurately gauge the metric scale of a 3D scene [9], [10], [11], [12], [13]. In contrast, combining radar and infrared sensors allows for delivering metric dense depth. Although the fusion of radar with RGB cameras for depth estimation has demonstrated considerable success [14], [15],

Manuscript received 2 February 2024; revised 29 May 2024; accepted 11 July 2024. Date of publication 5 August 2024; date of current version 1 November 2024. This work was supported by the National Natural Science Foundation of China under Grant U21A20484. The Associate Editor for this article was Y. Hou. *(Han Li and Yukai Ma are co-first authors.) (Corresponding authors: Xingxing Zuo; Yong Liu.)*

Han Li, Yukai Ma, Yuehao Huang, Yaqing Gu, Weihua Xu, and Yong Liu are with the Institute of Cyber-Systems and Control, Zhejiang University, Hangzhou 310058, China (e-mail: yongliu@iipc.zju.edu.cn).

Xingxing Zuo is with the School of Computation, Information and Technology, Technical University of Munich, 80333 Munich, Germany (e-mail: xingxing.zuo@tum.de).

Digital Object Identifier 10.1109/TITS.2024.3432996

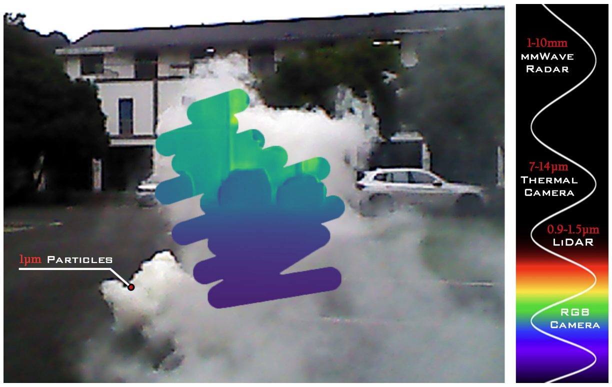


Fig. 1. Left: Our approach can provide high-quality depth estimation beyond the visible spectrum, unaffected by micrometer-sized particles. Right: Millimeter-wave radar and infrared thermal cameras have longer operational wavelengths than LiDAR and RGB cameras to penetrate atmospheric particles.

[16], [17], [18], the direct integration of either extracted features or raw data from infrared imagery with millimeter-wave radar data poses substantial challenges. These difficulties stem from the inherently low contrast and lack of textures in infrared images, combined with radar data's inherently noisy and sparse nature. Severe aliasing, blurring, and artifacts can compromise the quality of the resulting dense depth maps.

To overcome these challenges, this paper introduces RIDERS, a novel approach for dense depth estimation through the fusion of radar and thermal camera data. This work extends our previous conference paper RadarCam-Depth [19], which focused on combining radar with an RGB camera for depth estimation. Adopting a similar underlying philosophy, RIDERS is tailored to integrate radar and thermal camera data effectively, particularly in environments with smoke, fog, or low light. To the best of our knowledge, our work represents the ***first*** attempt to integrate an infrared camera and a radar for metric dense depth estimation. This synergistic sensor fusion yields significant advantages: (i) The integration of a thermal camera, which captures detailed relative dense depth, with radar, which provides metric scales for sampled points, enables the generation of metric dense depth maps with enhanced detail. (ii) Operating in the mid-infrared and millimeter-wave spectrums, these sensors effectively circumvent most atmospheric particulates, ensuring perception characterized by low diffraction and minimal occlusion. (iii) Immune to ambient light variations, thermal imaging, and millimeter-wave reflections facilitate consistent environmental perception, regardless of the time of day.

Specifically, we achieve the dense metric depth estimation from a radar and a thermal camera with three stages. Firstly, we employ an off-the-shelf generalizable monocular prediction model to obtain scaleless depth from the thermal images and perform global alignment with sparse radar point clouds. Concurrently, we learn the confidence of association between radar points and pixels in thermal images, enhancing sparse radar points into quasi-dense depth. Ultimately, we utilize the scale map learner to adjust local scales for monocular estimation results and recover the final dense depth. Our method tries to fully leverage the advantages of multi-modal data at each component stage. This three-stage paradigm boasts two significant advantages: (i) We circumvent the direct fusion of raw data or encodings of heterogeneous point clouds and images, thereby preventing aliasing artifacts and preserving high-fidelity fine details in dense depth estimation (see Fig. 1). (ii) Unlike direct depth estimation with a wide convergence basin, our approach essentially involves recovering dense scale for scale-free monocular depth, leveraging the strong prior of mono-depth and achieving higher learning efficiency.

The primary contributions of this work can be summarized as follows:

- Presenting the ***first*** known dense depth estimation approach that integrates mmWave radar and thermal cameras, possessing unparalleled robustness for depth perception in adverse conditions such as smoke and low lighting.
- Introducing a novel metric dense depth estimation framework that effectively fuses heterogeneous radar and thermal data. Our three-stage framework comprises monocular estimation and global alignment, quasi-dense radar augmentation, and dense scale learning, ultimately recovering dense depth from sparse and noisy long-wave data.
- The proposed method has undergone extensive testing on the publicly available NTU4DRadLM dataset [20] and the self-collected ZJU-Multispectrum dataset, surpassing other solutions and demonstrating high metric accuracy and solid robustness.
- Our high-quality ZJU-Multispectrum dataset containing challenging scenarios with 4D radar, thermal camera, RGB camera data, and reference depth from 3D LiDAR will be released. Both the dataset and our code will be released at https://github.com/MMOCKING/RIDERS to fertilize future research.

II. Related Works

A. Depth From Monocular Infrared Thermal Image

The infrared spectrum band exhibits high-level robustness against adverse weather and lighting conditions. However, infrared images lack texture information compared to visible spectrum images, appearing more blurred and suffering from a scarcity of large-scale datasets. Consequently, numerous existing methods attempt to transfer knowledge from the visible spectrum to thermal depth estimation tasks. Kim et al.'s Multispectral Transfer Network (MTN) [9] is trained with chromaticity clues from RGB images, enabling stable depth prediction from monocular thermal images. Lu and Lu [10] propose using a CycleGAN-based generator to transform RGB images into fake thermal images, creating a stereo pair of a thermal camera for supervising the disparity prediction. Shin et al. approach [11] leverages multispectral consistency for self-supervised depth estimation, incorporating temperature consistency from thermal imaging and photometric consistency from wrapped RGB images.

However, the above methods require closely matching and often RGB and thermal images from identical scenes and viewpoints, imposing stringent data requirements. Recently, Shin et al. [13], [21] proposed a method that does not require paired multispectral data. Their network

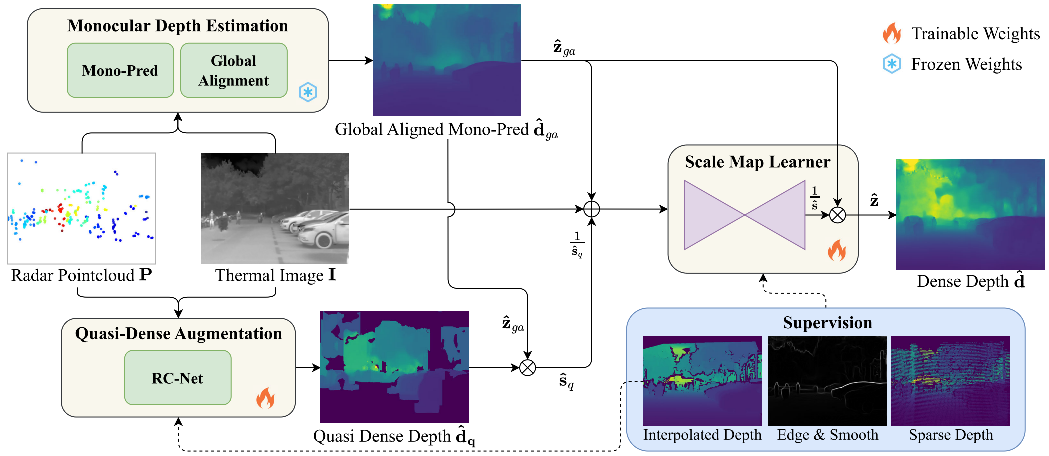

Monocular Depth Estimation
Mono-Pred
Global Alignment
Global Aligned Mono-Pred $\hat{\mathbf{d}}_{ga}$
$\hat{\mathbf{z}}_{ga}$
Trainable Weights
Frozen Weights
Radar Pointcloud $\mathbf{P}$
Thermal Image $\mathbf{I}$
$\frac{1}{\hat{\mathbf{s}}_q}$
Scale Map Learner
$\frac{1}{\hat{\mathbf{s}}}$
$\hat{\mathbf{z}}$
Dense Depth $\hat{\mathbf{d}}$
Quasi-Dense Augmentation
RC-Net
Quasi Dense Depth $\hat{\mathbf{d}}_{\mathbf{q}}$
$\hat{\mathbf{z}}_{ga}$
$\hat{\mathbf{s}}_q$
Supervision
Interpolated Depth
Edge & Smooth
Sparse Depth

Fig. 2. The overall framework of our proposed RIDERS is comprised of three stages: monocular depth estimation from infrared images, quasi-dense augmentation of radar depth, and scale map learner for refining the local scale of dense depth. $\mathbf{d}$ and $\mathbf{s}$ denote the depth and scale, while $\mathbf{z} = 1/\mathbf{d}$ is the inverse depth. The symbols $\oplus$ and $\otimes$ represent concatenation and element-wise multiplication, respectively.

consists of modality-specific feature extractors and modality-independent decoders. They train the network to achieve feature-level adversarial adaptation, minimizing the gap between RGB and thermal features. ThermalMonoDepth [12] is a self-supervised depth estimation method that eliminates the need for extra RGB involvement in training. It introduces a time-consistent image mapping method reorganizing thermal radiation values and ensuring temporal consistency, maximizing self-supervision for thermal image depth estimation. Additionally, Shin et al. [22] propose a unified depth network that effectively bridges monocular thermal depth and stereo thermal depth tasks from a conditional random field approach perspective. Nevertheless, monocular methods often lack accurate scale and are prone to local optima in self-supervised training, leading to poor metric accuracy.

B. Depth From Radar-Camera Fusion

The fusion of radar and RGB camera data for metric depth estimation is an active area of research. Lin et al. [14] introduced a two-stage CNN-based pipeline that combines radar and camera inputs to denoise radar signals and estimate dense depth. Long et al. [23] proposed a radar-2-Pixel (R2P) network, utilizing radial Doppler velocity and induced optical flow from images to associate radar points with corresponding pixel regions, enabling the synthesis of full-velocity information. They also achieved image-guided depth completion using radar and video data [15]. Another approach, DORN [16], proposed by Lo et al., extends radar points in the elevation dimension and applies a deep ordinal regression network-based [24] feature fusion. Unlike other methods, R4dyn [17] creatively incorporates radar as a weakly supervised signal into a self-supervised framework and employs radar as an additional input to enhance robustness. However, their method primarily focuses on vehicle targets and does not fully correlate all radar points with a larger image area, resulting in lower depth accuracy. Nevertheless, the methods above typically require multi-frame information to overcome the sparsity of radar data. In contrast, Singh et al. [18] presented a fusion method that relies solely on a single image frame and radar point cloud. Their first-stage network infers the confidence scores of associating a radar point to image patches, leading to a semi-dense depth map after association. They further employ a gated fusion network to control the fusion of multi-modal Radar-Camera data and predict the final dense depth.

These methods directly encode the multi-modal inputs and learn the target depth. However, direct encoding and concatenation of the inherently ambiguous radar depth and images can confuse the learning pipeline, resulting in aliasing and other undesirable artifacts in the estimated depth. In our preceding conference version [19], we explored scale learning for monocular depth estimation using RGB imagery supplemented by radar data. However, given the RGB camera's sensitivity to lighting conditions, there is a pressing need for a robust method of metric dense depth estimation using infrared cameras, which are unaffected by variations in ambient illumination. Compared to [19], we have enhanced the performance of radar depth augmentation by adjusting the quasi-dense depth calculation strategy. Additionally, the introduction of original thermal images as input, along with the incorporation of smoothness loss, has improved the performance of our Scale Map Learner.

III. Methodology

Our goal is to recover the dense depth $\hat{\mathbf{d}} \in \mathbb{R}_{+}^{H_0 \times W_0}$ from a pair of thermal image $\mathbf{I} \in \mathbb{R}^{C \times H_0 \times W_0}$ and radar point cloud $\mathbf{P} = \left\{\mathbf{p}_i | \mathbf{p}_i \in \mathbb{R}^3, i = 0, 1, 2, \cdots, k-1\right\}$ transformed into the thermal camera coordinate through known extrinsic calibration. As different datasets may use either single-channel thermal images or three-channel pseudo-color images, the number of channels C for the image $\mathbf{I}$ can be either 1 or 3. The overall framework of our RIDERS consists of three main stages: monocular depth prediction and alignment (Sec. III-A), quasi-dense augmentation of sparse radar (Sec. III-B), and scale map learner (SML) for refining dense scale locally (Sec. III-C), as shown in Fig. 2.

A. Monocular Depth Prediction and Scale Alignment

1) Monocular Depth Prediction: This module provides prior monocular depth $\hat{\mathbf{d}}_m$ for dense scale learning from a single-view thermal image. $\hat{\mathbf{d}}_m$ does not require high metric

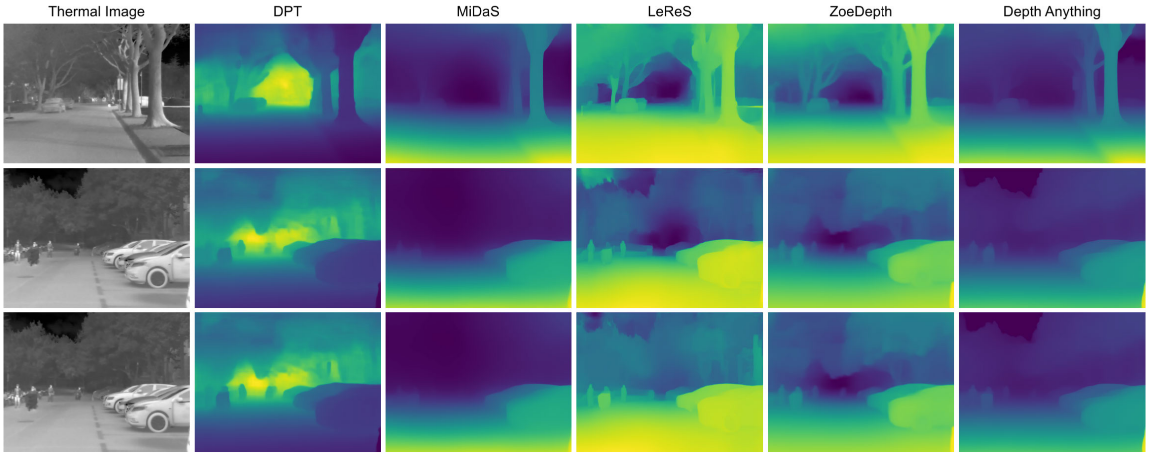


Fig. 3. **Zero-shot generalized depth predictions**. From left to right: the input thermal image, zero-shot generalized depth prediction from DPT [25], MiDaS [26], ZoeDepth [27], LeReS [28], [29], and Depth Anything [30], which are trained on RGB images. The second and third rows correspond to consecutive frames with a time interval of about 0.1 seconds. LeReS, ZoeDepth, and Depth Anything exhibit good performance without fine-tuning. Specifically, LeReS provides precise edges and fine details, while ZoeDepth and Depth Anything demonstrate temporal consistency in consecutive frames.

accuracy but needs to provide relative depth reflecting the high-fidelity image details, for example, object edges and surface smoothness.

Contrary to the abundance of large-scale datasets for RGB images in the visible spectrum, there is a notable scarcity of equivalent datasets for training monocular depth prediction models specifically for infrared images. However, thanks to the analogous imaging principles—both being passive receivers of electromagnetic waves—and the comparable modality between thermal and RGB images, we can directly utilize a monocular depth prediction model originally trained on RGB images for thermal depth prediction. This zero-shot generalization approach allows us to obtain preliminary dense depth for monocular thermal images.

With abundant datasets for monocular depth prediction for RGB images, monocular depth prediction designed for high generalization has gained increasingly robust performance through extensive training [25], [26], [27], [28], [29], [30], [31]. Existing methods rely on the depth reconstruction loss in the scale- and shift-invariant space on multiple datasets, enabling the recovery of high-quality scale-invariant relative depth. These models exhibit powerful zero-shot generalization capabilities, even when applied to monochrome thermal images. We can obtain the preliminary monocular depth $\hat{\mathbf{d}}_m$ (or inverse depth $\hat{\mathbf{z}}_m$) from the input $\mathbf{I}$ using any monocular depth prediction network. In other words, our module in this section utilizes a replaceable network, which can be continuously updated with the advancement of the monocular depth prediction networks. It is not necessary to retrain or fine-tune this network, as our subsequent trained Scale Map Learner (see Sec. III-C) will utilize information from radar to recover the dense metric scale for the monocular depth prediction, and alleviate its prediction error.

In the context of generalization on thermal images, pre-trained models like LeReS [28], [29], ZoeDepth [27], and Depth Anything [30] demonstrate reasonable accuracy. LeReS provides dense depth maps with clear object edges and structures, though it may exhibit depth discontinuities between adjacent frames. On the other hand, ZoeDepth and Depth Anything achieve high temporal consistency in their depth estimates, with Depth Anything particularly distinguished by its superior ability to differentiate background elements, such as the sky. (See Fig. 3).

2) Global Scale Alignment: In order to enhance the efficiency of refining pixel-wise scale by SML in the proceeding stage, we align the scale-free monocular depth prediction $\hat{\mathbf{d}}_m$ with the depth of radar points $\mathbf{P}$ using a global scaling factor $\hat{s}_g$, which generates globally aligned depth $\hat{\mathbf{d}}_{ga}$. Empirically, our findings suggest that a single scaling factor is adequate for global alignment, diverging from existing methodologies that employ both scaling and shifting factors for this purpose [19], [32].

To optimize the monocular depth estimate $\hat{\mathbf{d}}_m$, we employ the bounded Brent numerical optimization algorithm [33], [34]. Our optimization objective is to find the optimal scaling factor $\hat{s}_g$ that minimizes the following loss function:

$$\mathcal{F}(\hat{s}_g, \hat{\mathbf{d}}_m, d(\mathbf{P}), \mathbf{M}_{\mathbf{P}}) = \mathbf{M}_{\mathbf{P}} \cdot |\hat{s}_g \cdot \hat{\mathbf{d}}_m - d(\mathbf{P})|, \qquad (1)$$

where $\hat{s}_g$ is the scaling factor to be optimized, $d(\cdot)$ returns the depth of radar points in image coordinates, and $\mathbf{M}_{\mathbf{P}} \in \{0, 1\}^{H \times W}$ is the valid position mask of radar points on the image. We define radar points with depths in the range of 0-100m as valid ones.

We utilize empirical values to set initial bounds for the optimization target $\hat{s}_g$. Ultimately, the Brent algorithm employs a combination of quadratic interpolation, the secant method, and bisection to compute the optimal solution within the predefined bounds iteratively. The globally aligned metric monocular depth is then calculated as $\hat{\mathbf{d}}_{ga} = \hat{s}_g \cdot \hat{\mathbf{d}}_m$. Its inverse depth $\hat{\mathbf{z}}_{ga}$ is subsequently fed into the scale map learner (SML).

B. Quasi-Dense Radar Augmentation

Due to inherent sparsity and noises in radar data, additional augmentation of radar depth is crucial before conducting scale learning. To densify the sparse radar depth obtained from projection, we exploit a transformer-based Radar-Camera data association network (shorthand RC-Net), which predicts the confidence of radar-pixel associations. The proposed network

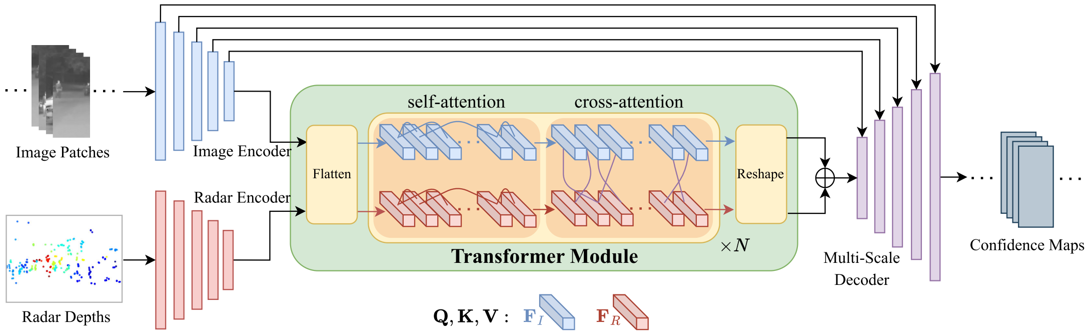


Fig. 4. **The architecture of our sparse radar augmentation network, RC-Net.** The network takes radar depths and image patches cropped around each radar point as input. The architecture consists of two encoder branches, a transformer module, and a multi-scale decoder, aiming to infer pixel-level confidence scores for each radar point-image patch pairing. The attention layers take query $\mathbf{Q}$, key $\mathbf{K}$ and value $\mathbf{V}$, from image features $\mathbf{F}_I$ and radar features $\mathbf{F}_R$.

predicts the association confidence between pixels in $\mathbf{I}$ and neighboring radar points. It utilizes a weighted average of the depths from multiple radar points to calculate the depth for each pixel. Ultimately, this module outputs a quasi-dense map with continuous depth, denoted as $\hat{\mathbf{d}}_q$.

1) Network Architecture: Our RC-Net (refer to Fig. 4) is derived from the vanilla RC-vNet [18] and is enhanced by integrating self and cross-attention mechanisms [35] within a transformer module. The image encoder of RC-Net adopts a standard ResNet backbone [36], while the radar encoder comprises a multi-layer perceptron with fully connected layers. We reproject the sparse radar points into the image plane to generate a sparse radar depth map, which is input into the radar encoder. The extracted radar features undergo mean pooling and are reshaped to match the width and height of image features. Subsequently, radar and image features are flattened and processed through $N = 4$ layers of self and cross-attention. To be specific, each attention layer in the transformer module takes inputs—query $\mathbf{Q}$, key $\mathbf{K}$ and value $\mathbf{V}$, from image features $\mathbf{F}_I$ and radar features $\mathbf{F}_R$. It calculates weights by assessing the similarity between $\mathbf{Q}$ and $\mathbf{K}$, converts similarity values to weights via a softmax layer, and produces the attention value by weighting and summing the values, as shown in the following equation:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V}. \tag{2}$$

Following the common practice, $\mathbf{K}$ and $\mathbf{V}$ are always from the same sensor modality, and d_k is the dimension of $\mathbf{K}$. The attention mechanism is self-attention when $\mathbf{Q}$ is from the same modality as $\mathbf{V}$, while it becomes across-attention when $\mathbf{Q}$ and $\mathbf{V}$ are from different modalities. Our transformer module draws inspiration from the linear attention mechanism design introduced in [35] and [37]. The attention operation intuitively selects relevant information by assessing the similarity between $\mathbf{Q}$ and each $\mathbf{V}$. It calculates the output vector as the weighted sum of the value vectors, where the similarity scores determine the weights. The projected radar depth and the image are fed into the transformer module, where self-attention and cross-attention encode and extract features of the multi-modal data in the image coordinate. This mechanism allows for better estimating the correlation between different modalities, providing our multi-modal fusion with a larger receptive field.

Ultimately, the features encoded through the transformer module, along with skip connections from the middle layers of the encoder, are input into the decoder. The output is in logit form, and the final step involves activating the logits through the sigmoid function, resulting in confidence maps for cross-modal associations.

2) Confidence of Cross-Modal Associations: For a radar point $\mathbf{p}_i$ and a cropped image patch $\mathbf{Z}_i \in \mathbb{R}^{C\times H\times W}$ in its projection vicinity, we utilize RC-Net h_θ to produce a confidence map $\hat{\mathbf{y}}_i = h_\theta(\mathbf{Z}_i, \mathbf{p}_i) \in [0, 1]^{H\times W}$, representing the probability of whether the pixels in $\mathbf{Z}_i$ corresponds to $\mathbf{p}_i$, inspired by [18]. With all k points in a radar point cloud $\mathbf{P}$, the forward pass generates k confidence maps for individual radar points. Therefore, each pixel $\mathbf{x}_{uv}$ within $\mathbf{I}$ ($u \in [0, W_0-1]$, $v \in [0, H_0-1]$) has $n \in [0, k]$ associated radar point candidates. By selecting confidence scores above the threshold, we can identify potential associated radar points $\mathbf{P}_\mu$ for pixel $\mathbf{x}_{uv}$. Then we compute the depth of pixel $\mathbf{x}_{uv}$ by taking a weighted average of all $\mathbf{P}_\mu$ depths using their normalized confidence scores as weights, resulting in a quasi-dense depth map $\hat{\mathbf{d}}_q$ as shown in Fig. 5.

$$\hat{\mathbf{d}}_q(u, v) = \begin{cases} \dfrac{\sum_{\mu\in\mathbb{U}} d(\mathbf{p}_\mu)\cdot\hat{\mathbf{y}}_\mu(x_{uv})}{\sum_{\mu\in\mathbb{U}}\hat{\mathbf{y}}_\mu(x_{uv})}, & \mathbb{U}\neq\emptyset, \\ \text{None}, & \text{otherwise}, \end{cases} \tag{3}$$

where $\mathbb{U} = \{i \mid \hat{\mathbf{y}}_i(x_{uv}) > \tau\}$, $\mathbf{p}_\mu$ is the associated radar point candidates, $\hat{\mathbf{y}}_\mu(x_{uv})$ is the corresponding confidence scores, $d(\cdot)$ returns the depth value. Finally, quasi-dense scale map $\hat{\mathbf{s}}_q$ is calculated from $\hat{\mathbf{s}}_q = \hat{\mathbf{d}}_q/\hat{\mathbf{d}}_{ga}$. Its inverse $1/\hat{\mathbf{s}}_q$ is subsequently fed into the scale map learner.

3) Training: We begin by projecting LiDAR point clouds to obtain $\mathbf{d}_{gt}$ in the image coordinate. Subsequently, we perform linear interpolation in log space [38] on $\mathbf{d}_{gt}$, resulting in $\mathbf{d}_{int}$. For supervision, we use $\mathbf{d}_{int}$ to create binary classification labels $\mathbf{y}_i \in \{0, 1\}^{H\times W}$, where $\mathbf{d}_{int}$ pixels with a depth value deviation less than 0.5m from the radar point are labeled as positive. Following the construction of $\mathbf{y}_i$, we minimize the binary cross-entropy loss:

$$\mathcal{L}_{BCE} = \frac{1}{|\Omega|}\sum_{x\in\Omega} -(\mathbf{y}_i(x)\log\hat{\mathbf{y}}_i(x) + (1-\mathbf{y}_i(x))\log(1-\hat{\mathbf{y}}_i(x))), \tag{4}$$

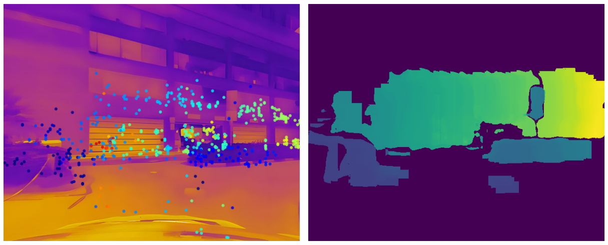


Fig. 5. **Radar augmentation result.** Left: sparse radar points back-projected onto the thermal image plane. Right: augmented quasi-dense depth $\hat{\mathbf{d}}_q$ from our RC-Net.

where $\Omega \subset \mathbb{R}^2$ denotes the image region of $\mathbf{Z}_i$, $x \in \Omega$ is a pixel coordinate, and $\hat{\mathbf{y}}_i = h_\theta(\mathbf{Z}_i, \mathbf{p}_i)$ is the confidence of correspondence.

C. Scale Map Learner

1) Network Architecture: Drawing inspiration from [44], we construct a scale map learner (SML) network based on MiDaS-small [31] architecture. SML aims to learn a pixel-wise dense scale map for $\hat{\mathbf{z}}_{ga}$, thereby completing the quasi-dense scale map and refining the metric accuracy of $\hat{\mathbf{z}}_{ga} = 1/\hat{\mathbf{d}}_{ga}$. SML requires concatenated $\mathbf{I}$, $\hat{\mathbf{z}}_{ga}$ and $1/\hat{\mathbf{s}}_q$ as input. The empty parts in $\hat{\mathbf{s}}_q$ are filled with ones. SML regresses a dense scale residual map $\mathbf{r}$, where values can be negative. The final scale map is derived as $1/\hat{\mathbf{s}} = \mathrm{ReLU}(1 + \mathbf{r})$, and the ultimate metric depth estimation is computed as $\hat{\mathbf{d}} = \hat{\mathbf{s}}/\hat{\mathbf{z}}_{ga}$.

2) Training: During training, ground truth depth $\mathbf{d}_{gt}$ is derived from the projection of 3D LiDAR points. Linear interpolation [38] in log space is further performed to get a densified $\mathbf{d}_{int}$. We minimize the difference between the estimated metric depth $\hat{\mathbf{d}}$ and the ground truth $\mathbf{d}_{gt}$ and $\mathbf{d}_{int}$ with a L1 penalty:

$$\mathcal{L}_{depth} = \mathcal{L}(\mathbf{d}_{int}, \hat{\mathbf{d}}) + \lambda_{gt}\mathcal{L}(\mathbf{d}_{gt}, \hat{\mathbf{d}}), \tag{5}$$

$$\mathcal{L}(\mathbf{d}, \hat{\mathbf{d}}) = \frac{1}{|\Omega_d|}\sum_{x \in \Omega_d} |\mathbf{d}(x) - \hat{\mathbf{d}}(x)|, \tag{6}$$

where λ_{gt} is the weight of $\mathcal{L}_{gt}$, $\Omega_d \subset \Omega$ denotes the domains where ground truth has valid depth values.

In addition, we incorporate a smoothness loss constraint. Leveraging the fact that $\hat{\mathbf{d}}_{ga}$ provides depth rather than texture-based edges, we exploit this to encourage smoothness in the non-edge regions of the output $\hat{\mathbf{d}}$, thereby enhancing the overall estimation quality. We employ Sobel filters [45] to compute the gradients $\triangledown x_{ga}$ and $\triangledown y_{ga}$ of $\hat{\mathbf{d}}_{ga}$, as well as the gradients $\triangledown x$ and $\triangledown y$ of $\hat{\mathbf{d}}$. This process allows us to formulate the smoothness loss $\mathcal{L}_{smooth}$:

$$\mathcal{L}_{smooth} = e^{-|\triangledown x_{ga}|} \cdot |\triangledown x| + e^{-|\triangledown y_{ga}|} \cdot |\triangledown y|. \tag{7}$$

Our final loss is expressed as:

$$\mathcal{L}_{SML} = \mathcal{L}_{depth} + \lambda_{smooth} \cdot \mathcal{L}_{smooth}, \tag{8}$$

where λ_{smooth} is the weight of $\mathcal{L}_{smooth}$.

IV. Experiments

A. Datasets

1) NTU4DRadLM Dataset: We first evaluate the proposed method on the NTU4DRadLM dataset [20]. Designed explicitly for SLAM and sensing research, this dataset integrates 4D radar, thermal cameras, and IMUs. The dataset is around 17.6km, 85 mins, and 50GB in total. To encompass both high-speed autonomous driving scenarios and low-speed robot scenarios, the dataset is divided into two parts: vehicle-mounted and handheld. The "loop" sequence represents data collected by vehicle-mounted sensors, while "cp", "garden", and "nyl" are sequences collected by handheld sensors.

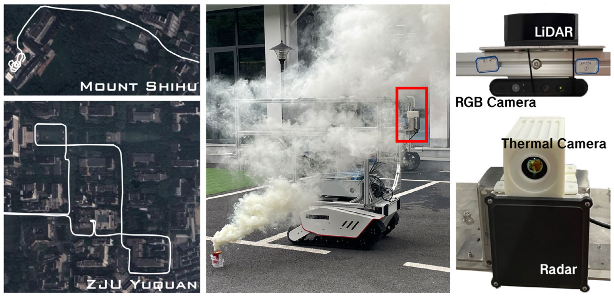


Fig. 6. **ZJU-Multispectrum dataset.** From left to right: data collection paths of ZJU-Multispectrum in Mount Shihu and Zhejiang University's Yuquan Campus, our vehicle equipped with multi-modal sensors, and the sensors mounted on the vehicle.

For testing our proposed RIDERS in common driving environments, specifically structured road scenes, we selected the "loop" sequence. Among these sequences, "loop2-2022-06-03-1" and "loop3-2022-06-03-0," consisting of a total of 6,039 time-synchronized Thermal-LiDAR-Radar keyframes, were used as the test set, while the remaining 24,160 frames of the "loop" sequence were utilized for training and validation. Additionally, to assess the generalization capability of our method in unstructured scenes, we applied the weights trained on the "loop" sequence to perform inference testing on the "garden" sequence, which consists of 6,876 frames. We synchronized the frames by pairing the thermal and radar data with the nearest timestamp to the lowest-frequency LiDAR frame. Moreover, point cloud projection is conducted according to the official extrinsics. It is worth noting that in the provided ROS topics for radar data, we exclusively used the "/radar_pcl" topic and did not utilize "/radar_enhanced_pcl".

However, it is important to highlight that the publicly available portion of the NTU4DRadLM dataset predominantly features typical scenes under clear weather conditions, which may limit the robustness assessment scope.

2) ZJU-Multispectrum Dataset: To assess the robustness of the proposed method in challenging conditions, such as smoke and nighttime environments, we collect data using our customized ground robot platform (see Fig. 6), proposing our ZJU-Multispectrum dataset.

The comparison between ZJU-Multispectrum and existing similar datasets is shown in Tab. I. Our dataset comprises high-quality 4D imaging radar data (where 4D represents spatial dimensions xyz and Doppler dimension) along with multispectral RGB/thermal data, providing scenes for testing robustness in both darkness and heavy smoke, a feature absent in all previous datasets. Although RRxIO [7] also provides radar and thermal data, it is based on a low-power platform mounted on drones, limiting its scene coverage. In contrast, our dataset is primarily tailored for road scenes in autonomous driving, aligning more closely with practical development needs. Additionally, our dataset utilizes solid-state MEMS

TABLE I

Comparison of Publicly Available Radar Datasets

Dataset	Radar Type	LiDAR Type	Thermal Data	Adverse Scenes	Scale
nuScenes [39]	3D	32-beam mechanical	×	dark, rainy	large
RadarScenes [40]	3D	×	×	rainy	large
Astyx [41]	4D	16-beam mechanical	×	×	large
RRxIO [7]	4D	×	✓	dark, foggy	small
VoD [42]	4D	64-beam mechanical	×	×	large
TJ4DRadSet [43]	4D	32-beam mechanical	×	dark	large
NTU4DRadLM [20]	4D	non-repetitive scanning	✓	dark	large
ZJU-Multispectrum	**4D**	**solid-state MEMS**	✓	**dark, heavy smoke**	**large**

(Micro-Electro-Mechanical Systems) LiDAR to generate high-quality and dense ground truth, offering practical and precise supervision and evaluation for our methods. Compared to mechanical LiDAR, front-view solid-state LiDAR is better suited for our depth estimation task.

Our data collection vehicle (see Fig. 6) is equipped with the following sensors: Geometrical-pal 4D imaging radar, RoboSense M1 LiDAR, Dali VD641 thermal infrared camera, and Orbbec Astra RGBD camera. Utilizing these sensors, we collected data across various scenarios, encompassing daytime, nighttime, clear weather, and conditions with artificial smoke generated by smoke canisters. The multi-sensor calibration follows the principle of aligning to the LiDAR. We first calibrate the transformation relationship of the camera and LiDAR and then perform point cloud registration from the radar point cloud to the LiDAR point cloud. The frame rates of the sensors we used are LiDAR 10 Hz, radar 13 Hz, and thermal/RGB camera 30 Hz. Due to the high frame rates of the sensors and the slow movement speed of our data collection vehicle (around 5 km/h), we directly retrieve the nearest timestamp from the LiDAR frame with the lowest frame rate to obtain synchronized frames for multi-modal data.

Our training and validation sets involve sequences captured in clear weather, consisting of eight daytime sequences and one nighttime sequence, 21,522 synchronized frames in total. For quantitative evaluation in daytime scenarios, we selected three clear-weather sequences comprising a total of 5,532 frames. To assess performance in more challenging conditions, we utilized three additional sequences. Specifically, we examined the method's resilience to low-light environments through a nighttime clear-weather sequence of 184 frames. Furthermore, we evaluated the method's effectiveness in smoke conditions with one daytime and one nighttime smoke sequence, together providing LiDAR depth ground truth for 446 frames. It should be noted that the sensors were maintained in a stationary position for the 446 frames with depth ground truth. This was essential to ensure the acquisition of LiDAR depth of the surroundings, which served as the reference for our evaluation. The remaining frames of two smoke sequences, with sensors moving in the floating and diffusing smoke, are used for the visualization showcase.

B. Training Details and Evaluation Protocol

For both two datasets, LiDAR point clouds projected onto the image plane were used as the ground truth $\mathbf{d}_{gt}$ for training and evaluation. Subsequently, linear interpolation [38] was performed on $\mathbf{d}_{gt}$ in the logarithmic space of depth, yielding $\mathbf{d}_{int}$ as the dense supervisory signal.

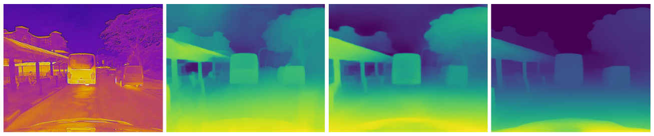


Fig. 7. **Mono-Preds of the NTU4DRadLM images.** From left to right: the input thermal image, scale-free inv-depth $\hat{\mathbf{z}}_m$ from LeReS [28], ZoeDepth [27] and Depth Anything [30].

For training the RC-Net for quasi-dense radar augmentation, we adopted the network architecture of Sec. III-B. When training on NTU4DRadLM, with an input image size of 512 × 640, the size of the cropped patch for confidence map generation in RC-Net is set to 150 × 50. For ZJU-Multispectrum, the input image size is 480×640, and the patch size is 240 × 100. The training employed the Adam optimizer, with $\beta_1 = 0.9$ and $\beta_2 = 0.999$, and a learning rate of $2e^{-4}$ for 100 epochs. Data augmentations, including horizontal flipping, saturation, brightness, and contrast adjustments, are applied with a probability of 0.5.

We employ the MiDaS-Small network architecture for our Scale Map Learner (SML). The encoder backbone is initialized with pre-trained ImageNet weights [46], and other layers are randomly initialized. The input data is resized to a fixed height of 288 and a width that is a multiple of 32. We use an Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The initial learning rate is set to $1e^{-4}$ and reduced to $5e^{-5}$ after 20 epochs. Data augmentations, including horizontal flipping and random radar noise, are employed. All training and testing activities were carried out on a single RTX 3090 GPU.

Some widely adopted metrics from the literature are used for evaluating the depth estimations, including mean absolute error (MAE), root mean squared error (RMSE), absolute relative error (AbsRel), squared relative error (SqRel), the errors of inverse depth (iRMSE, iMAE), and δ_1 [47]. To enhance clarity in our presentation, depth evaluation metrics are calculated in units of mm, and inverse depth metrics are calculated in units of km^{-1}. To better illustrate our experimental details, we have prepared a demo video, which is available for viewing at https://youtu.be/wRsRTZoWUpE.

C. Evaluation on NTU4DRadLM

1) Structured Road Scenarios: This section evaluates the metric dense depth against $\mathbf{d}_{gt}$ within the range of 50, 60, and 70 meters on the "loop" sequences of NTU4DRadLM [20] dataset. Due to the lack of existing methods based on

TABLE II

Evaluations on NTU4DRadLM (Structured Road)

Range	Method	iMAE ↓	iRMSE ↓	MAE ↓	RMSE ↓	AbsRel ↓	SqRel ↓	δ_1 ↑
0-50m	Depth Anything [30]	6.164	11.151	**1688.480**	**3583.927**	0.095	**787.197**	0.905
	DORN [16]	5.903	10.575	1915.141	4458.021	0.090	806.773	0.920
	Singh [18]	5.340	10.256	1866.829	4586.367	0.092	1092.425	0.915
	RacarCam-Depth [19]	5.065	10.173	2004.375	4869.808	0.094	1157.161	0.908
	RIDERS (LeReS)	4.585	9.539	1785.232	4552.015	0.084	1029.812	0.921
	RIDERS (ZoeDepth)	**4.377**	9.459	1745.794	4552.851	**0.082**	1053.554	0.923
	RIDERS (Depth Anything)	4.496	**9.442**	1752.455	4465.857	**0.082**	1001.022	**0.924**
0-60m	Depth Anything [30]	6.075	11.037	**1949.512**	**4137.690**	0.098	**868.378**	0.898
	DORN [16]	6.040	10.664	2323.629	5329.139	0.097	979.009	0.908
	Singh [18]	5.307	10.270	2120.584	5076.233	0.094	1154.673	0.909
	RacarCam-Depth [19]	5.091	10.293	2321.156	5493.308	0.097	1246.593	0.900
	RIDERS (LeReS)	4.590	9.633	2040.465	5056.552	0.086	1094.782	0.915
	RIDERS (ZoeDepth)	**4.381**	9.544	1988.879	5027.688	**0.084**	1111.248	**0.918**
	RIDERS (Depth Anything)	4.496	**9.526**	1996.693	4947.669	0.085	1060.204	**0.918**
0-70m	Depth Anything [30]	5.982	10.937	**2177.917**	**4642.064**	0.099	**926.953**	0.894
	DORN [16]	6.112	10.706	2622.396	6013.591	0.101	1096.660	0.900
	Singh [18]	5.288	10.289	2344.816	5577.079	0.096	1222.454	0.904
	RacarCam-Depth [19]	5.122	10.427	2619.444	6152.787	0.100	1349.044	0.892
	RIDERS (LeReS)	4.602	9.713	2283.267	5606.387	0.088	1171.620	0.909
	RIDERS (ZoeDepth)	**4.392**	9.619	2217.410	5538.507	**0.086**	1179.719	**0.913**
	RIDERS (Depth Anything)	4.508	**9.609**	2230.930	5484.166	0.087	1133.040	**0.913**

Radar-Infrared camera fusion, we compare with methods [16], [18], [19] originally designed for radar-RGB fusion. We also fine-tuned and evaluated SOTA monocular approach, Depth Anything [30], to highlight the advantages of our fusion pipeline. All methods were thoroughly trained and validated on the thermal images and radar point cloud data from the dataset. We also tested the RIDERS method using different monocular depth estimation modules, including LeReS [28], ZoeDepth [27], and Depth Anything [30].

Compared to RIDERS with monocular estimates from LeReS, RIDERS using ZoeDepth and Depth Anything exhibit superior metric accuracy. This is attributed to the excellent consistency in their predictions, resulting in fewer depth discontinuities between consecutive frames. This characteristic is advantageous for precise scale learning in the SML module. Compared to the fine-tuned Depth Anything [30], our RIDERS significantly outperforms it in iMAE and iRMSE. Given that radar measurements are denser and more accurate at close ranges, our fusion method excels in near-range depth estimation, which is more critical in perception. At farther distances, where radar points are sparse with larger errors, depth inference mainly relies on the image. Consequently, RIDERS performs similarly to the fine-tuned Depth Anything at long ranges. Moreover, the decreasing inverse depth error of Depth Anything with increasing range indicates its lower proficiency in estimating depths at close distances.

On the other hand, our method also outperforms other Radar-Camera fusion methods on most evaluation metrics (see Tab. II). Within the evaluation ranges of 50m, 60m, and 70m, RIDERS (ZoeDepth) shows a reduction of 13.6%, 13.9%, and 14.3% in iMAE compared to the second-best method, respectively. Similarly, AbsRel is reduced by 8.9%, 10.6%, and 10.4%, respectively. We attribute the outstanding performance of RIDERS to the reasonable monocular depth prediction $\hat{\mathbf{d}}_m$ (see Fig. 7) and our scale learning strategy. RIDERS exhibits excellent qualitative results regarding the edges and morphology of objects such as vehicles and buildings, as shown in Fig. 8. Scale learning based on $\hat{\mathbf{d}}_{ga}$ outputs avoids issues like blurring and deformation that may arise in direct-depth metric learning methods [16], [18]. Compared to our previous work, RadarCam-Depth [19] developed for RGB images, RIDERS has effectively improved the adaptability and accuracy for low-contrast and blurry thermal images. The augmented radar depth calculation in our RC-Net greatly enhances the metric accuracy (see Sec. IV-E.3). Furthermore, the introduction of raw thermal imagery as an input to the SML provides an auxiliary source of information, thereby refining the predictions derived solely from $\hat{\mathbf{d}}_{ga}$.

2) Unstructured Garden Scenarios: Recently, methods for depth inference using only a single image have seen rapid development. For example, Depth Anything [30], based on its weights trained on mixed datasets, requires only a small amount of fine-tuning on a new dataset to achieve good metric depth. However, when the testing scenario differs significantly from the training scenario, the metric accuracy of monocular depth estimation methods significantly degrades. In contrast, our method utilizes additional radar measurements to obtain metric depth, providing an advantage when generalizing to entirely new scenes.

To demonstrate the superiority of our proposed RIDERS over monocular methods, we conducted tests in the unstructured scenes of the "garden" sequence. Following the splits of the "loop" sequence described in the previous section, we fine-tuned Depth Anything using data from the the training set of structured road scenes. Similarly, our RIDERS directly used weights trained on the "loop" sequence for inference on the "garden" sequence.

The scenes in the "garden" sequences are filled with extensive vegetation, have a disordered temperature distribution, and

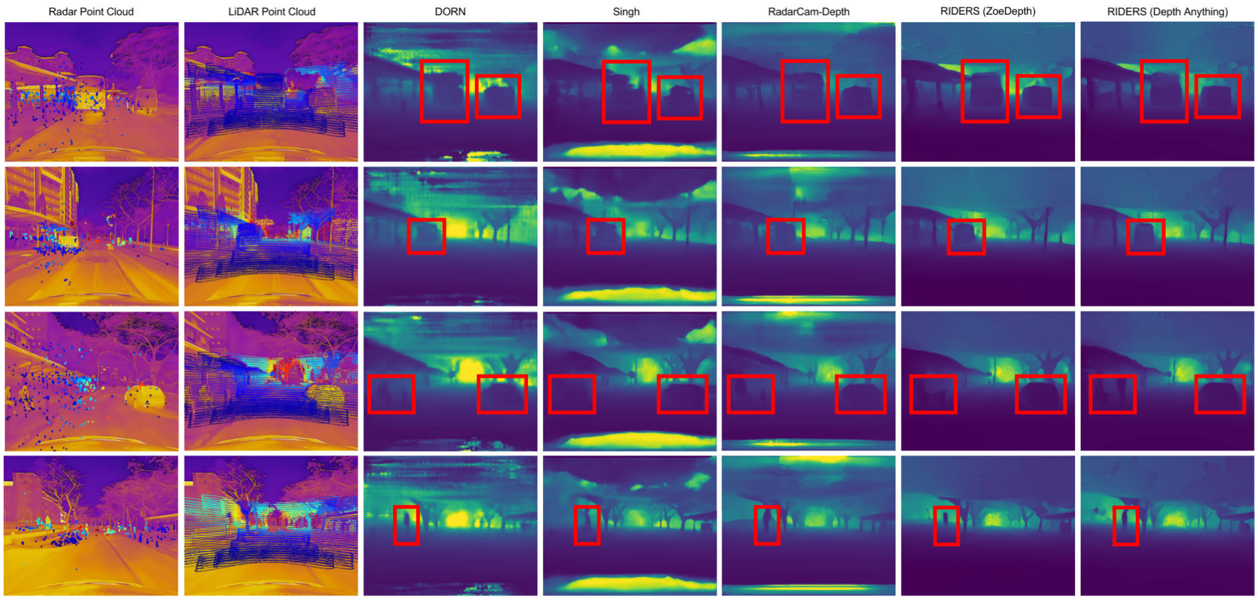


Fig. 8. **Evaluations on NTU4DRadLM.** From left to right: Raw radar points $\mathbf{P}$ overlaid on the infrared image, LiDAR point clouds $\hat{\mathbf{d}}_{gt}$ overlaid on the infrared image, and depth predictions from $\hat{\mathbf{d}}$ of DORN [16], Singh [18], and our RIDERS based on monocular depth prediction models, ZoeDepth [27] and Depth Anything [30], on NTU4DRadLM dataset [20]. Harnessing the preliminary monocular depth prediction, our method demonstrates superior performance. Specifically, for crucial objects in traffic scenes, such as vehicles and pedestrians, our approach provides clear outlines of the targets (highlighted by red boxes in the images). The monocular depth predictions from the pre-trained ZoeDepth and Depth Anything models, corresponding to the first row, are illustrated in Fig. 7.

TABLE III

EVALUATIONS ON NTU4DRADLM (UNSTRUCTURED GARDEN)

Range	Method	iMAE ↓	iRMSE ↓	MAE ↓	RMSE ↓	AbsRel ↓	SqRel ↓	δ_1 ↑
0-50m	Depth Anything [30]	69.393	89.925	8790.493	11499.380	0.949	12174.003	0.124
	RIDERS (Depth Anything)	**64.805**	**85.765**	**7846.082**	**10844.823**	**0.834**	**10349.292**	**0.152**

TABLE IV

EVALUATIONS ON ZJU-MULTISPECTRUM (CLEAR-DAY)

Range	Method	iMAE ↓	iRMSE ↓	MAE ↓	RMSE ↓	AbsRel ↓	SqRel ↓	δ_1 ↑
0-50m	DORN [16]	11.184	19.176	2975.231	5376.028	0.162	1256.015	0.786
	Singh [18]	11.282	17.386	2979.062	5579.791	0.164	1547.239	0.788
	RacarCam-Depth [19]	10.647	16.484	2917.770	5433.830	0.167	1611.953	0.798
	Direct Depth (ZoeDepth)	11.763	17.231	2923.381	4822.382	0.170	1175.563	0.763
	Sparse Input (ZoeDepth)	10.035	15.149	2721.244	5019.822	0.148	1219.950	0.806
	RIDERS (LeReS)	10.423	15.755	2658.893	4765.552	0.147	1084.529	0.806
	RIDERS (ZoeDepth)	9.413	14.291	2576.012	4755.345	0.137	1081.774	0.826
	RIDERS (Depth Anything)	**8.495**	**13.796**	**2423.499**	**4587.857**	**0.127**	**1025.969**	**0.848**
0-60m	DORN [16]	11.255	20.526	3213.565	5983.413	0.166	1376.385	0.779
	Singh [18]	11.292	17.420	3247.038	6099.891	0.167	1664.067	0.781
	RacarCam-Depth [19]	10.587	16.433	3115.653	5772.308	0.168	1668.136	0.793
	Direct Depth (ZoeDepth)	11.745	17.201	3213.202	5422.726	0.173	1288.683	0.753
	Sparse Input (ZoeDepth)	10.011	15.130	2947.856	5483.022	0.150	1303.773	0.798
	RIDERS (LeReS)	10.443	15.788	2939.058	5363.344	0.150	1202.025	0.797
	RIDERS (ZoeDepth)	9.435	14.341	2830.197	5297.403	0.140	1188.412	0.818
	RIDERS (Depth Anything)	**8.510**	**13.822**	**2655.641**	**5076.447**	**0.130**	**1114.202**	**0.840**
0-70m	DORN [16]	11.287	20.546	3345.805	6394.996	0.167	1453.923	0.775
	Singh [18]	11.298	17.444	3384.523	6407.994	0.168	1726.581	0.777
	RacarCam-Depth [19]	10.575	16.422	3231.842	6023.848	0.169	1711.887	0.791
	Direct Depth (ZoeDepth)	11.772	17.220	3419.844	5945.049	0.175	1390.536	0.747
	Sparse Input (ZoeDepth)	10.029	15.155	3112.811	5878.983	0.152	1380.064	0.794
	RIDERS (LeReS)	10.482	15.835	3128.167	5843.109	0.152	1297.107	0.792
	RIDERS (ZoeDepth)	9.461	14.379	2992.761	5699.685	0.142	1266.466	0.814
	RIDERS (Depth Anything)	**8.532**	**13.852**	**2806.616**	**5449.972**	**0.131**	**1182.343**	**0.836**

exhibit complex structures. Consequently, the details in the infrared images are very blurry, making accurate depth estimation challenging. Moreover, due to the significant differences from the training set, both our method and Depth Anything experience noticeable accuracy degradation when generalizing. However, as shown in Tab. III, our method still maintains

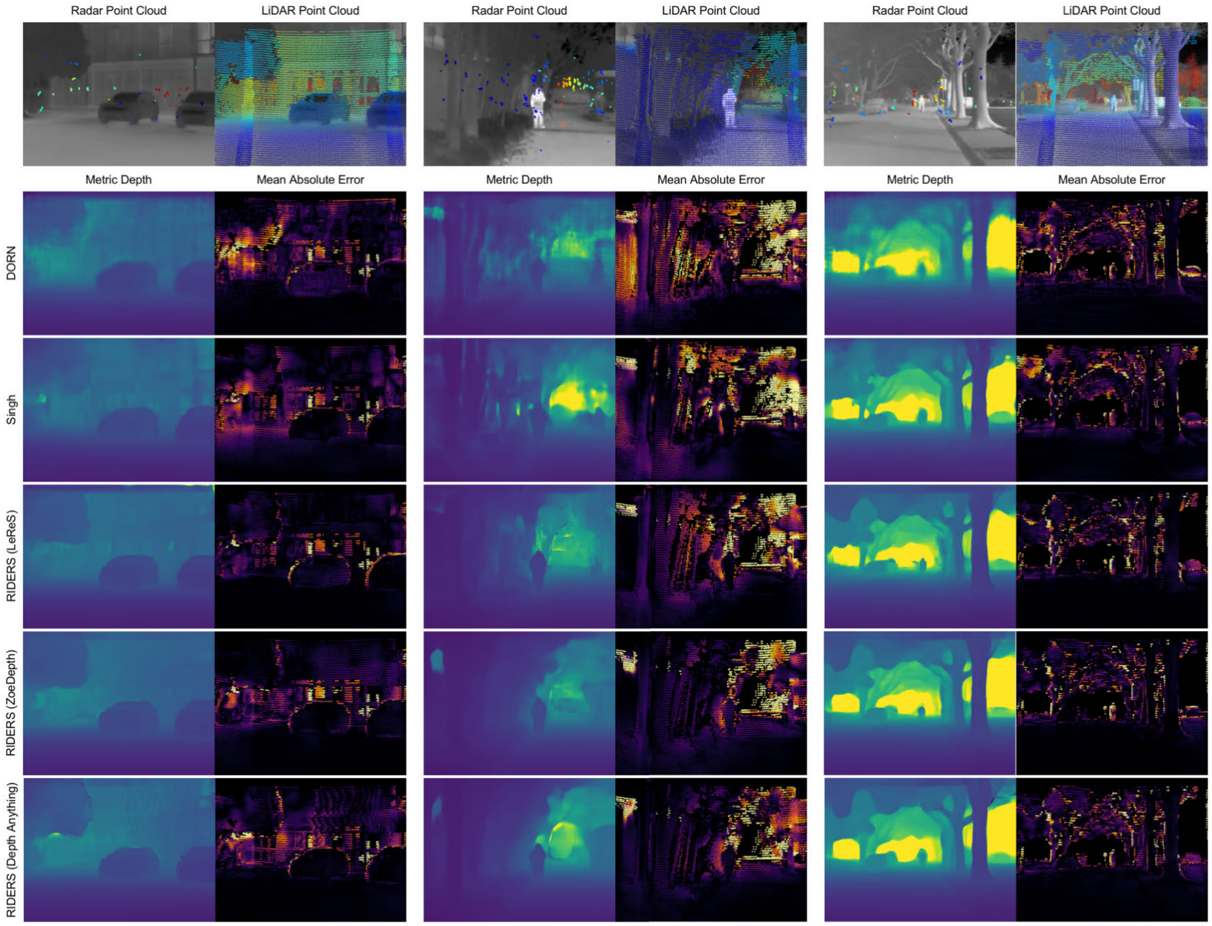


Fig. 9. **Evaluations on ZJU-Multispectrum (clear-day).** First row: radar (left) and LiDAR (right) point cloud of ZJU-Multispectrum dataset. The remaining five rows, from top to bottom, show the depth estimation results (left) of DORN [16], Singh [18], and our RIDERS ($\hat{\mathbf{d}}_m$ from LeRes [28] and ZoeDepth [27]) along with error visualizations (right).

a clear advantage over Depth Anything, outperforming the monocular method on all metrics.

D. Evaluation on ZJU-Multispectrum

1) Clear Daytime: We follow a similar way to Sec. IV-C for the evaluations on the ZJU-Multispectrum dataset. We first evaluate RIDERS on three clear daytime sequences. Our RIDERS outperformed DORN [16], Singh [18] and RadarCam-Depth [19] across all metrics, as shown in Tab. IV and Fig. 9. Regarding monocular depth estimation modules, Depth Anything is the best choice. Within the evaluation ranges of 50m, 60m, and 70m, RIDERS (Depth Anything) exhibits a reduction in iMAE compared to the second-best method by 20.2%, 19.6%, and 19.3%, respectively. In terms of iRMSE, RIDERS (Depth Anything) achieves a reduction of 16.3%, 15.9%, and 15.6% compared to the second-best method within the same ranges. Our method also demonstrates good performance on AbsRel, with the error reduced by 21.6%, 21.7%, and 21.6% at 50m, 60m, and 70m, respectively, compared to the second-best method.

2) Nighttime or Smoke Scenarios: The primary objective of this work is to address the issue of perception (depth estimation) failure in adverse conditions. In this section, we evaluate our method in typical adverse scenarios under low-light and smoke conditions. For safety reasons, simulations of adverse scenarios were conducted in enclosed areas, resulting in a limited range of depth values. The evaluation of metric accuracy is calculated within a range of 50 meters. Results are shown in Tab. VI.

To evaluate the performance of our method under different lighting conditions, we conducted tests on both a clear-day sequence and a low-light sequence captured in the same scene. Notably, in clear-night conditions, our RIDERS method, benefiting from the low-light insensitivity of both the thermal camera and radar, showcased accuracy comparable to that observed in the clear-day sequence.

To simulate challenging smoke-laden scenarios, we created artificial smoke with about 1 μm particle diameters. After that, we conducted tests when the sensor suite was kept stationary or moving in the presence of smoke. In sequences affected by smoke, LiDAR measurements are proved to be unreliable due to interference from the smoke particles. To assess metric depth estimation under smoky conditions, we maintained a stationary position for the sensor suite. We then used LiDAR depths from previously captured stationary frames (absent of smoke) as a reference for evaluating frames that were impacted by smoke (see Fig. 10). In this case, 446 synchronized radar and thermal frames were available for metric evaluation. Our RIDERS exhibits high accuracy in both the nighttime

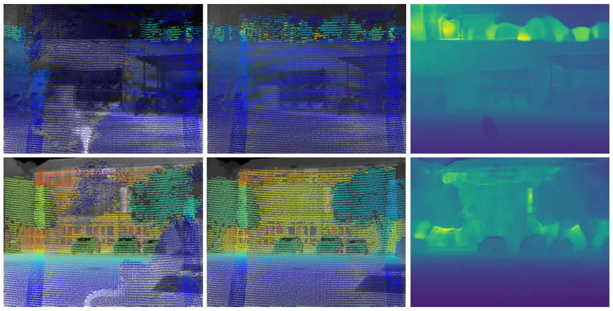


Fig. 10. **Showcased results on the smoke sequences of ZJU-Multispectrum dataset.** From left to right: LiDAR depth with smoke, its corresponding ground truth depth from the historical frame without smoke, and the depth estimation from RIDERS.

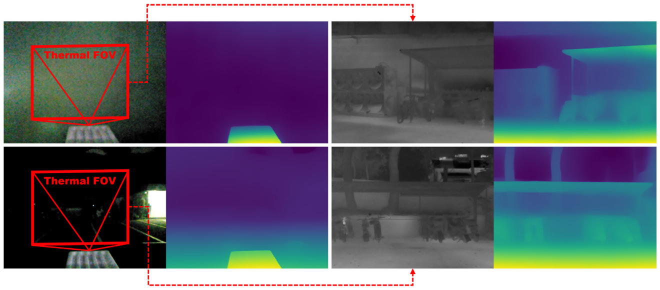


Fig. 11. **Comparison of RGB and thermal cameras.** From left to right: the RGB images with their Depth Anything [30] predictions, and the thermal images with their Depth Anything [30] predictions. Each row points to the same scene at the same time. The leftmost RGB images show the thermal camera with its FOV.

and daytime sequences with smoke, shown in Tab. VI, and outperforms the other comparative methods [16], [18], [19]. To conduct thorough evaluations, we also subjected RIDERS to tests while the sensor suite was in motion, focusing on qualitative assessments across two smoke-impacted sequences. These tests further confirmed the robustness of RIDERS in handling smoke and nighttime conditions, as illustrated in Fig. 12.

As our experiments demonstrated, the negligible impact of low light and atmospheric particulates on thermal cameras and millimeter-wave radars ensures that RIDERS maintains its efficacy in scenarios severely compromising short-wave sensors such as LiDAR and RGB cameras. Consequently, RIDERS can deliver dependable metric depth estimates with consistent reliability across diverse conditions, including smoke-laden environments and both daytime and nighttime settings.

It is essential to highlight that, compared to RGB camera-based solutions, our thermal camera offers unparalleled advantages in nighttime and smoke scenarios. Thermal cameras rely on the infrared radiation emitted by objects rather than the reflection of external ambient light, enabling them to work in the darkness. Moreover, due to its longer wavelength, infrared light experiences less diffraction in the atmosphere, which allows it to penetrate smoke. Therefore, as shown in Fig. 11, our thermal camera can provide clear scene images even in adverse conditions. We employ Depth Anything [30] to infer scale-free depth using the images captured by RGB and thermal cameras. The experiments demonstrate that thermal images enable adequate depth perception, whereas RGB images often result in blurred and inaccurate depth information under challenging conditions. Although the two cameras' differing focal lengths and FOVs (Field of View) make it difficult to compare their monocular depth prediction accuracy fairly quantitatively, the visualizations in specific scenarios clearly highlight the hardware advantages of the thermal camera.

TABLE V
MODULES RUNTIME OF RIDERS (S)

Mono-Pred [30]	Global Alignment	RC-Net	SML
0.0415	0.0088	0.0793	0.0299

3) Run Time: To assess the efficiency of the proposed method in this paper, we conducted a statistical analysis of the runtime of each module in RIDERS (including data loading and processing and model inference). To mitigate bias caused by data distribution, we used all 39,142 frames from the ZJU-Multispectrum dataset for the runtime calculation, taking the average computation time per frame. The input thermal image size is 640×480, and each frame of the radar point cloud contains 163 radar points on average. When using RC-Net to compute the correlation between radar points and surrounding pixel neighborhoods, the size of the neighborhood patch is 240×100. The final results are shown in Tab. V. Among these, Mono-Pred & Global Alignment can run in parallel with RC-Net. Ideally, the total processing time for a single frame is around 0.1 seconds.

E. Ablation

Unless additional clarification is provided, the ablation studies referenced in this paper are conducted using the clear daytime sequences from the ZJU-Multispectrum dataset.

1) Local Scale Refinement: To validate the efficacy of our Scale Map Learner module, we evaluated the performance of the $\hat{\mathbf{d}}_{ga}$ without undergoing the local scale refinement described in Sec. III-C. As shown in Tab. VII, zero-shot monocular depth prediction from LeReS [28], ZoeDepth [27], and Depth Anything [30] after global scale alignment with sparse radar depth still exhibit poor accuracy, highlighting the need for further pixel-wise scale learning. Despite the lower metric accuracy of the $\hat{\mathbf{d}}_{ga}$ from Depth Anything, the inconsistency between local and global scales can be corrected through our SML. Our RIDERS (Depth Anything) demonstrates the best performance overall compared to the comparisons.

2) Scale Learning Strategy: We executed a series of ablation experiments to examine the proposed approach of learning metric scale for monocular depth estimation. To provide a comparison, we trained the scale map learner (SML) with metric depth supervision, coercing it to output depth directly instead of scale. Upon testing, directly learning depth resulted in blurry depth output and reduced convergence efficiency, as illustrated in Fig. 13. The final evaluation results are also in the "Direct Depth" rows of Tab. IV. Notably, the strategy of directly learning depth produces significantly lower accuracy compared to learning metric scale for monocular depth.

These facts are attributed to the inherent characteristics of the sensors we use. The SML network takes input from monocular depth prediction, original thermal images, and

TABLE VI
EVALUATIONS ON ZJU-MULTISPECTRUM (NIGHT/SMOKE)

Sequence	Method	iMAE ↓	iRMSE ↓	MAE ↓	RMSE ↓	AbsRel ↓	SqRel ↓	δ_1 ↑
Clear-Day	DORN [16]	7.626	10.707	2498.583	4001.246	0.122	699.280	0.838
	Singh [18]	9.592	14.019	2955.268	4755.387	0.148	1105.797	0.781
	RacarCam-Depth [19]	9.558	13.431	2983.785	4748.783	0.159	1115.849	0.782
	RIDERS (LeReS)	7.987	11.428	2473.828	4005.093	0.122	701.728	0.835
	RIDERS (ZoeDepth)	**5.973**	**8.730**	2081.379	3519.996	**0.098**	526.128	0.897
	RIDERS (Depth Anything)	6.123	8.903	**2006.024**	**3441.091**	0.099	**499.517**	**0.899**
Clear-Night	DORN [16]	9.117	11.994	2273.446	3825.647	0.121	645.963	0.794
	Singh [18]	8.986	11.796	2228.139	3852.676	0.124	653.333	0.832
	RacarCam-Depth [19]	8.119	11.081	2166.027	3785.561	0.125	673.940	0.847
	RIDERS (LeReS)	8.244	11.309	2108.396	3950.610	0.109	621.577	0.848
	RIDERS (ZoeDepth)	7.873	10.796	2082.876	3785.653	0.109	598.189	0.832
	RIDERS (Depth Anything)	**6.752**	**9.242**	**1705.355**	**3198.993**	**0.093**	**440.929**	**0.895**
Smoke-Night	DORN [16]	3.947	5.941	1300.994	2522.735	0.067	272.588	0.954
	Singh [18]	3.900	6.589	1273.503	2697.400	0.066	299.175	0.952
	RacarCam-Depth [19]	4.948	7.692	1836.872	3269.745	0.095	496.367	0.915
	RIDERS (LeReS)	**3.143**	**5.129**	1123.309	2430.145	**0.054**	219.441	0.965
	RIDERS (ZoeDepth)	3.635	5.633	**1110.664**	**2273.283**	0.056	**188.195**	**0.969**
	RIDERS (Depth Anything)	3.530	5.848	1175.626	2402.630	0.057	215.350	0.962
Smoke-Day	DORN [16]	6.901	9.991	4424.448	7120.837	0.149	1472.820	0.788
	Singh [18]	5.870	8.712	4539.365	7798.484	0.149	1796.066	0.792
	RacarCam-Depth [19]	5.736	6.792	4048.047	6615.171	0.139	1322.224	0.839
	RIDERS (LeReS)	**4.047**	5.545	**2702.623**	4091.246	**0.092**	513.222	0.901
	RIDERS (ZoeDepth)	4.182	**5.039**	2812.817	4711.455	0.095	616.680	0.924
	RIDERS (Depth Anything)	5.318	6.801	2765.294	**3866.913**	0.105	**486.345**	**0.926**

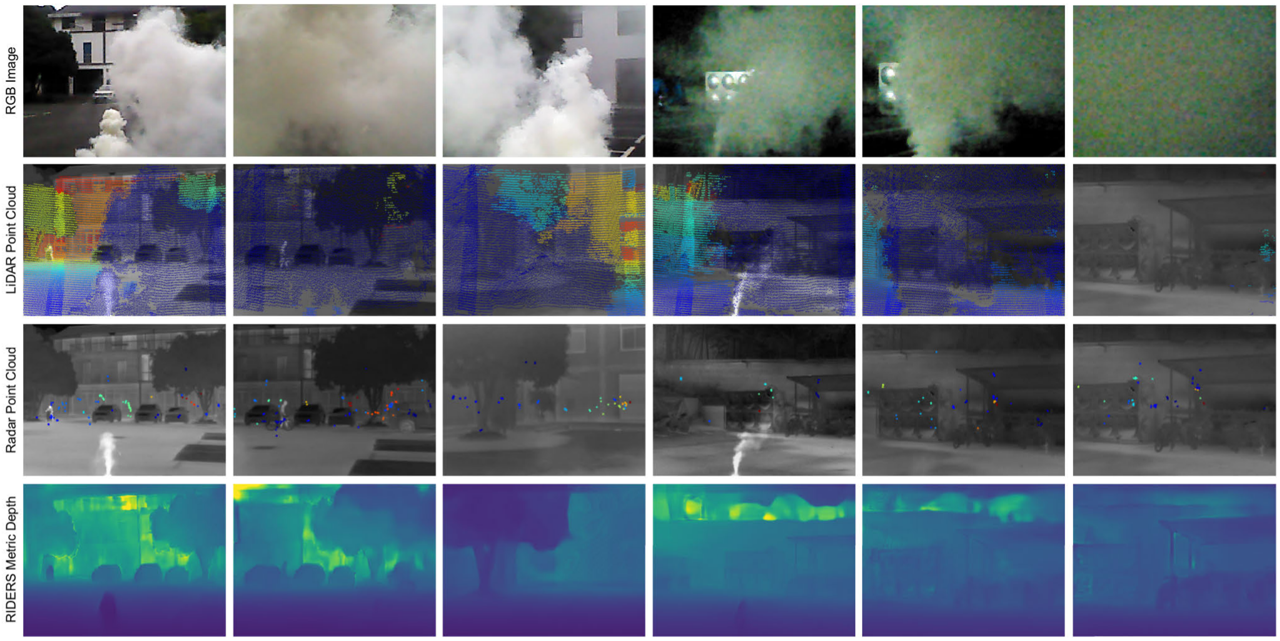


Fig. 12. **Evaluations on ZJU-Multispectrum (smoke).** From top to bottom: RGB images in the visible spectrum, LiDAR point clouds, radar point clouds, dense depth estimations of RIDERS. RIDERS performs remarkably well in scenarios where methods relying on visible light cameras and near-infrared LiDAR are severely impaired or face destructive challenges.

TABLE VII
ABLATION OF LOCAL SCALE REFINEMENT

Range	Method	iMAE ↓	iRMSE ↓	MAE ↓	RMSE ↓	AbsRel ↓	SqRel ↓	δ_1 ↑
0-50m	LeReS (Zero-Shot) [28]	26.420	32.966	**6788.466**	**12056.420**	**0.381**	**7651.739**	0.424
	ZoeDepth (Zero-Shot) [27]	**22.064**	**27.780**	9384.187	20058.585	0.488	31493.616	**0.494**
	Depth Anything (Zero-Shot) [30]	33.815	37.611	28224.456	50243.447	1.387	136253.932	0.112
	RIDERS (LeReS)	10.423	15.755	2658.893	4765.552	0.147	1084.529	0.806
	RIDERS (ZoeDepth)	9.413	14.291	2576.012	4755.345	0.137	1081.774	0.826
	RIDERS (Depth Anything)	**8.495**	**13.796**	**2423.499**	**4587.857**	**0.127**	**1025.969**	**0.848**

augmented radar depth. Due to the sparsity and measurement noise of the radar, pixel misalignment inevitably occurs when concatenating monocular depth prediction, original thermal images, and augmented radar depth directly. This

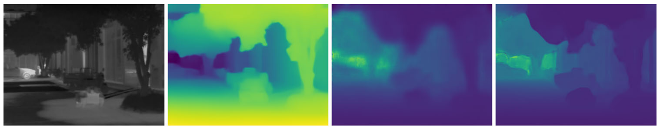


Fig. 13. **Ablation of the scale learning strategy.** From left to right: the input thermal image, monocular depth estimation from ZoeDepth [27], metric depth estimation obtained through direct depth learning, and metric depth estimation obtained through scale learning.

TABLE VIII

ABLATION OF RADAR AUGMENTATION

Method	iMAE ↓	iRMSE ↓	MAE ↓	RMSE ↓	Output Pts ↑
RC-vNet [18]	9.796	15.389	1191.473	2457.033	26692.283
RC-vNet (Avg) [18]	7.251	12.003	991.886	2288.220	26692.283
RC-Net (Avg)	**6.289**	**9.820**	**921.780**	**1948.600**	**26791.745**

misalignment existing in the input of SML is eventually manifested as blurriness of the output depth estimation. If a network directly predicts depth maps with large propagated gradients during the training, this blurriness will be reflected as unclear object edges. However, if the network outputs a dense scale map with smaller gradients that represents local refinement, and the output will be multiplied onto $\hat{\mathbf{d}}_{ga}$, the impact of defective feature association blurriness and unclear edges will be significantly reduced. In other words, the preliminary depth from monocular prediction should not be viewed solely as a feature channel for learning. However, it should directly provide its most fundamental meaning – depth, to our final depth estimation. Even if this preliminary monocular depth prediction does not have absolute metric accuracy, it can serve as a valid base to be explicitly corrected by the predicted dense scale map from SML.

3) Radar Depth Augmentation: We conducted ablation experiments to analyze the role of RC-Net for radar depth augmentation in the overall framework. As a comparison, we directly input the unenhanced sparse radar points into SML for scale learning. The final dense depth estimation results are shown in the "Sparse Input" rows in Tab. IV, lagging behind in all metrics compared to the quasi-dense radar input. Before the augmentation of RC-Net, only a few hundred radar points were projected onto the image plane per frame. The augmented quasi-dense depth could provide a more extensive range of metric scales for SML. The "Output Pts" in Tab. VIII denotes the number of points (pixels) with valid depth after radar depth augmentation.

4) RC-Net Internal Components: To validate the effectiveness of our improvements to RC-vNet (see Sec. III-B), we also test the benefits of the transformer module with attention mechanism and the quasi-dense depth calculation strategy on the NTU4DRadLM dataset. Our RC-Net, compared to the RC-vNet, incorporates a transformer module into its network architecture. In terms of quasi-dense depth calculation, RC-vNet assigns the radar depth with the highest confidence to the pixel, resulting in a blocky distribution and discontinuous depth changes in the augmented depth, deviating from the actual physical scene. In contrast, our method "RC-Net (Avg)" employs a weighted average of the depths of multiple radar points, making the quasi-dense depth closer to a continuous distribution in the real world, as shown in Fig. 14. Indeed, due to noise in radar points and uncertainties in confidence inference, aggregating information from multiple points provides more context information, benefiting the depth augmentation.

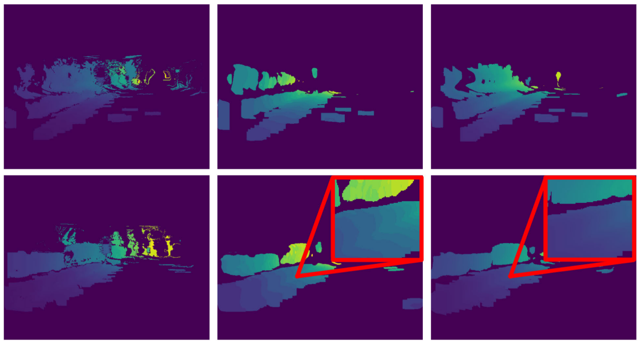


Fig. 14. **Ablation of radar depth augmentation.** From left to right: the visualizations from cross-modal associations labels, the augmented depth from RC-vNet [18], and the augmented depth from our proposed RC-Net. The detailed views of the quasi-dense depth are shown in the red box, demonstrating that our RC-Net can provide coherent and smooth depth estimation.

We evaluated quasi-dense depth $\hat{\mathbf{d}}_q$ in the range of 0-50m using the interpolated dense ground truth $\mathbf{d}_{int}$. As shown in Tab. VIII, the network architecture and the quasi-dense depth calculation strategy we employed significantly improved the accuracy of the augmented radar depth. To independently validate the effectiveness of the transformer module, we modified the calculation strategy of RC-vNet to a weighted average, as shown in the "RC-vNet (Avg)" row of Tab. VIII. It still falls behind our proposed RC-Net.

V. CONCLUSION

This paper introduces a novel method for robust metric depth estimation, combining millimeter-wave radar point clouds with infrared thermal images. To address the performance degradation of existing depth estimation methods in challenging scenarios such as nighttime and smoke, we employed long-wave, low-light-impact radar, and infrared thermal cameras as sensors to achieve robust depth estimation. We complement the metric depth estimation with a three-stage fusion process, which addresses the blurriness and significant noises for long-wave electromagnetic inputs. This process includes monocular depth prediction and global alignment, sparse radar depth augmentation, and local scale refinement. Each stage leverages the respective advantages of multi-modal inputs effectively. Our approach based on learning local scale refinement avoids artifacts, noise, and blurring caused by the direct fusion of heterogeneous data encodings, showcasing good qualitative and quantitative results. In summary, we integrate a radar and an infrared thermal camera for metric depth estimation in adverse scenarios, which addresses the existing void in Radar-Infrared fusion-based depth estimation. We also proposed the challenging ZJU-Multispectrum dataset and thoroughly tested our algorithm across diverse scenarios from multiple datasets. The conclusive findings of our study underscore that our pioneering multi-modal data fusion approach markedly improves upon the accuracy of current methodologies.

However, our proposed method still has several limitations. The reference depths obtained from LiDAR, used for supervisory purposes, cover only a limited portion of the image, potentially complicating the learning-based scale refinement process. Therefore, our future research may focus on introducing more prosperous supervisory signals based on modalities such as images while also exploring the possibility of self-supervision. Furthermore, our approach significantly depends on preliminary monocular depth prediction, and the adaptability of models pre-trained on RGB datasets to thermal infrared imagery necessitates further enhancement. While we have replaced the fine-tuning process of pre-trained models with the Scale Map Learner module, we should explore more data augmentation and preprocessing methods to address the differences between thermal and RGB images. Lastly, our method consists of multiple stages, leaving room for efficiency and resource consumption improvement.

References

[1] C. Rablau, "LiDAR—A new (self-driving) vehicle for introducing optics to broader engineering and non-engineering audiences," in *Education and Training in Optics and Photonics*. Washington, DC, USA: Optica Publishing Group, 2019.

[2] X. Zhao, L. Liu, R. Zheng, W. Ye, and Y. Liu, "A robust stereo feature-aided semi-direct SLAM system," *Robot. Auto. Syst.*, vol. 132, Oct. 2020, Art. no. 103597.

[3] L. Liu, X. Song, M. Wang, Y. Liu, and L. Zhang, "Self-supervised monocular depth estimation for all day images using domain separation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 12737–12746.

[4] S. Gasperini, N. Morbitzer, H. Jung, N. Navab, and F. Tombari, "Robust monocular depth estimation under challenging conditions," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 8177–8186.

[5] A. Tessmann, S. Kudszus, T. Feltgen, M. Riessle, C. Sklarczyk, and W. H. Haydl, "Compact single-chip W-band FMCW radar modules for commercial high-resolution sensor applications," *IEEE Trans. Microwave Theory Techn.*, vol. 50, no. 12, pp. 2995–3001, Dec. 2002.

[6] W. Zhangu, Z. Jun, D. Chuanguang, G. Xin, and Y. Kai, "Traffic vehicle cognition in severe weather based on radar and infrared thermal camera fusion," *Meas. Sci. Technol.*, vol. 32, no. 9, Sep. 2021, Art. no. 095111.

[7] C. Doer and G. F. Trommer, "Radar visual inertial odometry and radar thermal inertial odometry: Robust navigation even in challenging visual conditions," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Sep. 2021, pp. 331–338.

[8] J. Zhang et al., "4DRT-SLAM: Robust SLAM in smoke environments using 4D radar and thermal camera based on dense deep learnt features," in *Proc. IEEE Int. Conf. Cybern. Intell. Syst. (CIS) IEEE Conf. Robot., Autom. Mechatronics (RAM)*, Jun. 2023, pp. 1–20.

[9] N. Kim, Y. Choi, S. Hwang, and I. S. Kweon, "Multispectral transfer network: Unsupervised depth estimation for all-day vision," in *Proc. AAAI Conf. Artif. Intell.*, 2018, vol. 32, no. 1, pp. 1–26.

[10] Y. Lu and G. Lu, "An alternative of LiDAR in nighttime: Unsupervised depth estimation based on single thermal image," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 3832–3842.

[11] U. Shin, K. Lee, S. Lee, and I. S. Kweon, "Self-supervised depth and ego-motion estimation for monocular thermal video using multi-spectral consistency loss," *IEEE Robot. Autom. Lett.*, vol. 7, no. 2, pp. 1103–1110, Apr. 2022.

[12] U. Shin, K. Lee, B.-U. Lee, and I. S. Kweon, "Maximizing self-supervision from thermal image for effective self-supervised learning of depth and ego-motion," *IEEE Robot. Autom. Lett.*, vol. 7, no. 3, pp. 7771–7778, Jul. 2022.

[13] U. Shin, K. Park, B.-U. Lee, K. Lee, and I. S. Kweon, "Self-supervised monocular depth estimation from thermal images via adversarial multi-spectral adaptation," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2023, pp. 5787–5796.

[14] J.-T. Lin, D. Dai, and L. V. Gool, "Depth estimation from monocular images and sparse radar data," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2020, pp. 10233–10240.

[15] Y. Long, D. Morris, X. Liu, M. Castro, P. Chakravarty, and P. Narayanan, "Radar-camera pixel depth association for depth completion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 12502–12511.

[16] C.-C. Lo and P. Vandewalle, "Depth estimation from monocular images and sparse radar using deep ordinal regression network," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2021, pp. 3343–3347.

[17] S. Gasperini, P. Koch, V. Dallabetta, N. Navab, B. Busam, and F. Tombari, "R4Dyn: Exploring radar for self-supervised monocular depth estimation of dynamic scenes," in *Proc. Int. Conf. 3D Vis. (3DV)*, Dec. 2021, pp. 751–760.

[18] A. Deep Singh et al., "Depth estimation from camera image and mmWave radar point cloud," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 9275–9285.

[19] H. Li, Y. Ma, Y. Gu, K. Hu, Y. Liu, and X. Zuo, "Radarcam-depth: Radar-camera fusion for depth estimation with learned metric scale," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, Jun. 2024, pp. 1–11.

[20] J. Zhang et al., "NTU4DRadLM: 4D radar-centric multi-modal dataset for localization and mapping," in *Proc. IEEE 26th Int. Conf. Intell. Transp. Syst. (ITSC)*, Sep. 2023, pp. 4291–4296.

[21] U. Shin, K. Park, K. Lee, B.-U. Lee, and I. S. Kweon, "Joint self-supervised learning and adversarial adaptation for monocular depth estimation from thermal image," *Mach. Vis. Appl.*, vol. 34, no. 4, p. 55, Jul. 2023.

[22] U. Shin, J. Park, and I. S. Kweon, "Deep depth estimation from thermal image," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 1043–1053.

[23] Y. Long, D. Morris, X. Liu, M. Castro, P. Chakravarty, and P. Narayanan, "Full-velocity radar returns by radar-camera fusion," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 16178–16187.

[24] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao, "Deep ordinal regression network for monocular depth estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2002–2011.

[25] R. Ranftl, A. Bochkovskiy, and V. Koltun, "Vision transformers for dense prediction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 12179–12188.

[26] R. Birkl, D. Wofk, and M. Muller, "MiDaS v3.1—A model zoo for robust monocular relative depth estimation," 2023, *arXiv:2307.14460*.

[27] S. Farooq Bhat, R. Birkl, D. Wofk, P. Wonka, and M. Muller, "ZoeDepth: Zero-shot transfer by combining relative and metric depth," 2023, *arXiv:2302.12288*.

[28] W. Yin et al., "Learning to recover 3D scene shape from a single image," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 204–213.

[29] W. Yin et al., "Towards accurate reconstruction of 3D scene shape from a single monocular image," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 5, pp. 6480–6494, May 2023.

[30] L. Yang, B. Kang, Z. Huang, X. Xu, J. Feng, and H. Zhao, "Depth anything: Unleashing the power of large-scale unlabeled data," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Sep. 2024, pp. 10371–10381.

[31] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler, and V. Koltun, "Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 3, pp. 1623–1637, Mar. 2022.

[32] J. Naumann, B. Xu, S. Leutenegger, and X. Zuo, "NeRF-VO: Real-time sparse visual odometry with neural radiance fields," *IEEE Robot. Autom. Lett.*, vol. 9, no. 8, pp. 7278–7285, Aug. 2024.

[33] G. E. Forsythe et al., *Computer Methods for Mathematical Computations*. Upper Saddle River, NJ, USA: Prentice-Hall, 1977.

[34] R. P. Brent, *Algorithms for Minimization Without Derivatives*. Chelmsford, MA, USA: Courier Corporation, 2013.

[35] J. Sun, Z. Shen, Y. Wang, H. Bao, and X. Zhou, "LoFTR: Detector-free local feature matching with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 8922–8931.

[36] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[37] L. Li et al., "Geo-localization with transformer-based 2D-3D match network," *IEEE Robot. Autom. Lett.*, vol. 8, no. 8, pp. 4855–4862, Aug. 2023.

[38] C. B. Barber, D. P. Dobkin, and H. Huhdanpaa, "The quickhull algorithm for convex hulls," *ACM Trans. Math. Softw.*, vol. 22, no. 4, pp. 469–483, Dec. 1996.

[39] H. Caesar et al., "nuScenes: A multimodal dataset for autonomous driving," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11621–11631.

[40] O. Schumann et al., "Radarscenes: A real-world radar point cloud data set for automotive applications," in *Proc. IEEE Int. Conf. Inf. Fusion (FUSION)*, Nov. 2021, pp. 1–8.

[41] M. Meyer and G. Kuschk, "Automotive radar dataset for deep learning based 3D object detection," in *Proc. 16th Eur. Radar Conf. (EuRAD)*, 2019, pp. 129–132.

[42] A. Palffy, E. Pool, S. Baratam, J. F. P. Kooij, and D. M. Gavrila, "Multi-class road user detection with 3+1D radar in the view-of-delft dataset," *IEEE Robot. Autom. Lett.*, vol. 7, no. 2, pp. 4961–4968, Apr. 2022.

[43] L. Zheng et al., "TJ4DRadSet: A 4D radar dataset for autonomous driving," in *Proc. IEEE 25th Int. Conf. Intell. Transp. Syst. (ITSC)*, Oct. 2022, pp. 493–498.

[44] D. Wofk, R. Ranftl, M. Muller, and V. Koltun, "Monocular visual-inertial depth estimation," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2023, pp. 6095–6101.

[45] O. Vincent and O. Folorunso, "A descriptive algorithm for Sobel image edge detection," in *Proc. InSITE Conf.*, 2009, pp. 97–107.

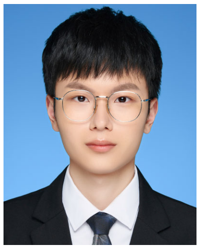
[46] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2009, pp. 248–255.

[47] J. Sun, Y. Xie, L. Chen, X. Zhou, and H. Bao, "NeuralRecon: Real-time coherent 3D reconstruction from monocular video," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 15598–15607.

Han Li received the B.Eng. degree in control science and engineering from Zhejiang University, Hangzhou, China, in 2022, where he is currently pursuing the master's degree with the College of Control Science and Engineering. His research interests include deep learning in sensor fusion, perception, and SLAM.

Yukai Ma received the B.Eng. degree in electrical engineering and its automation from Zhejiang University of Technology in 2021. He is currently pursuing the Ph.D. degree with the College of Control Science and Engineering, Zhejiang University, Hangzhou, China. He is a full-time Researcher with Shanghai Artificial Intelligence Laboratory, exploring the application of large models in autonomous driving. His research interests include deep learning in sensor fusion and SLAM.

Yuehao Huang received the B.Eng. degree in intelligence science and technology from Hangzhou Dianzi University in 2023. He is currently pursuing the Ph.D. degree with the College of Control Science and Engineering, Zhejiang University, Hangzhou, China. His research interests include SLAM, embodied intelligence, and 3D reconstruction.



Yaqing Gu received the B.Eng. degree in control science and engineering from Zhejiang University, Hangzhou, China, in 2023, where he is currently pursuing the master's degree with the College of Control Science and Engineering. His research interests include deep learning in sensor fusion and SLAM.

Weihua Xu received the Ph.D. degree in control theory and engineering from Zhejiang University in 2003. She is currently an Associate Professor with the Institute of Cyber-Systems and Control, College of Control Science and Engineering, Zhejiang University.

Yong Liu received the B.S. degree in computer science and engineering and the Ph.D. degree in computer science from Zhejiang University, Hangzhou, China, in 2001 and 2007, respectively. He is currently a Professor with the Institute of Cyber-Systems and Control, College of Control Science and Engineering, Zhejiang University. His research interests include machine learning, robotics vision, multiple-sensor fusion, and intelligent systems.

Xingxing Zuo received the B.Eng. degree in mechanical engineering from the University of Electronic Science and Technology of China, Chengdu, China, in 2016, and the Ph.D. degree in control science and engineering from Zhejiang University, Hangzhou, China, in 2021. He was a Post-Doctoral Researcher with the Technical University of Munich and a Research Faculty Researcher (Scientist) with Google. He is currently a Post-Doctoral Researcher with California Institute of Technology. His research interests include robotic perception and intelligence. He was the Finalist for the Best Paper Award in Robot Vision from ICRA 2021. He is an Associate Editor of IEEE ROBOTICS AND AUTOMATION LETTERS, IROS, and ICRA.