GLOBAL-LOCAL FEATURE ENHANCEMENT NETWORK FOR ROBUST OBJECT DETECTION USING MMWAVE RADAR AND CAMERA

Kaikai Deng, Dong Zhao, Qiaoyue Han, Zihan Zhang, Shuyue Wang, Huadong Ma

Beijing Key Lab of Intelligent Telecommunication Software and Multimedia,
Beijing University of Posts and Telecommunications, Beijing 100876, China

ABSTRACT

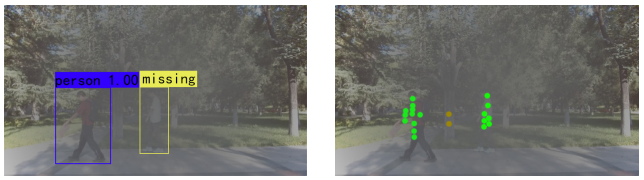
Object detection with camera has achieved promising results using deep learning methods, but it suffers degraded performance under adverse conditions (e.g., foggy weather, poor illumination). To remedy this, some recent studies resort to leveraging the complementary mmWave radar, which is less affected by adverse conditions, and designing effective fusion methods. However, the existing early fusion methods are vulnerable to data noise, while the existing late fusion methods ignore the association of object information between feature maps in the early stage. To overcome these shortcomings, we propose a Global-Local Feature Enhancement Network (GLE-Net), a two-stage deep fusion detector, which first generates anchors from two sensors and uses an auxiliary module to locally enhance the single-branch missing proposals, and then fuses the global features from the multimodal sensors to improve final detection results. We collect two datasets under foggy weather and poor illumination conditions with diverse scenes, and conduct extensive experiments, verifying that the proposed GLE-Net surpasses other state-of-the-art methods in terms of Average Precision (AP).

Index Terms— object detection, mmWave radar, auxiliary module, global-local fusion, deep learning

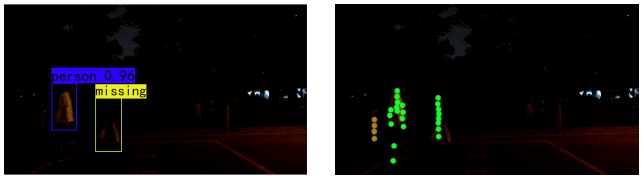
1. INTRODUCTION

Object detection is a fundamental computer vision task in various intelligent systems. The development of deep learning has promoted the application of object detection in many fields, such as search and rescue [1], farmland thief detection [2], and traffic monitoring [3]. Latest image-based methods [1, 4, 5, 6] have achieved promising results, but they suffer degraded performance under adverse conditions [7, 8]. Specifically, Figs. 1(a) and (b) show that the image detector, faster-rcnn [5], has missed detection under foggy weather and poor illumination conditions.

This work was supported in part by the National Key Research and Development Program of China under No. 2018AAA0101201, the NSFC under No. 61732017 and No. 61972044, the Fundamental Research Funds through the Central Universities under No. 2020XD-A09, the Funds for International Cooperation and Exchange of NSFC under No. 61720106007, and the 111 Project under No. B18008.



(a) Detection with camera (left) and mmWave radar (right) in foggy weather



(b) Detection with camera (left) and mmWave radar (right) in poor illumination

(c) Detection with camera (left) and mmWave radar (right) in normal condition

Fig. 1. Examples of object detection results using camera and mmWave radar. The camera can detect objects close to each other under normal condition, but fails under foggy weather and poor illumination conditions. Conversely, the mmWave radar can detect objects under foggy weather and poor illumination conditions, but fails to detect objects close to each other under normal condition.

Aside from camera, mmWave radar is regarded as an additional modality for autonomous sensing. Due to the ability to work under adverse conditions and the penetrability to airborne obstacles, mmWave radar is widely used in non-ideal conditions where camera is constrained. Figs. 1(a) and (b) show that the mmWave radar can detect persons under foggy weather and poor illumination conditions using DBSCAN [9]. Nevertheless, the point cloud generated by mmWave radar is sparse and noisy, resulting in a weaker ability of separating multiple objects than camera, as shown in Fig. 1(c). Neither the mmWave radar-based detector nor the camera-based detector alone performs satisfactorily in above scenes, which motivates the necessity of fusing mmWave radar and camera.

Many fusion methods have been proposed for camera-

radar setups. The early fusion performs detection by concatenating the image and mmWave radar features [10, 11], but it is vulnerable to data noise. To robustly detect objects, the late fusion methods separate the feature extraction network for each sensor stream, and concatenate the outputs of two sensor branches into a final multi-channel feature map or fuse their proposals [7, 8]. Such a late fusion scheme is able to generate sufficient feature maps or proposals as the fusion is focused on the region of interest (RoIs), but it ignores the association of object information between feature maps in the early stage. Moreover, existing fusion models usually need to be trained from scratch using multi-modality datasets. Unlike large-scale image datasets such as COCO [12] and PASCAL VOC [13], the available camera-mmWave radar datasets [14] are mainly used for highly specific systems and scenes. Therefore, users generally need to process and annotate their self-built datasets, which is labor-intensive and costly.

To address the above challenges, one should take advantage of both early fusion (rich data) and late fusion (focus on ROIs) while overcoming their shortcomings. To this end, we propose a Global-Local Feature Enhancement Network (GLE-Net) for robust object detection under foggy weather and poor illumination conditions. Specifically, GLE-Net consists of two stages. The first stage extracts the camera and mmWave radar feature maps through a *feature extractor*, and uses an *anchor generator* to obtain the initial anchors, followed by an *auxiliary module* to process the aggregated anchors to locally enhance the single-branch missing proposals. The second stage employs the global proposals generated by the two sensors' feature maps for fusion via a *variant transformer*. Besides, we adopt a loosely coupled fusion architecture that can be trained with less labeled multi-modality data. Specifically, the image branch is pre-trained with large-scale public datasets, and the mmWave radar branch is trained with a small amount of data as it is insensitive to the appearance of objects. Finally, we collect two datasets under diverse scenes and conduct extensive experiments. Experimental results show that GLE-Net improves the average precision (AP) by 14.5% and 2.9% compared with two state-of-the-art fusion methods, milliEye [7] and Naive Fusion [8], respectively, under foggy weather condition, and improves by 19.2% and 9.7% respectively under poor illumination condition.

2. RELATED WORK

Recent multi-modal methods [15, 16, 17] for object detection have shown that complementing camera images with depth and semantics has great potential to improve detection performance. RRPN [18] and RPR [19] use an early fusion detector, which projects the radar point clouds to 2D images and adopts a region proposal network (RPN) for joint detection. Deep Entropy Fusion [14] also utilizes an early fusion detector, which quantifies multi-sensor features with entropy to selectively fuse features. In addition, RadarNet [10] adopts the CNN to extract early features and further uses radar's radial velocity to refine motion prediction. LiRaNet [11] processes sparse radar points via a spatio-temporal feature extractor and fuses it with lidar data and road map to predict trajectories of objects. However, the early fusion often introduces data noise, resulting in false alarms and large regression errors.

To robustly detect objects, Chadwick et.al [8] design a late fusion detector, which feeds the radar and image data into the CNN for fusion via the concatenation operation. milliEye [7] decouples the image and mmWave radar branches for training separately and exploits the refinement module to increase the object classification threshold to ensure that the bounding boxes of objects can be kept. In contrast, we aggregate the anchors generated by the two branches to achieve local enhancement and use the generated global proposals to further optimize the fusion network.

3. PROPOSED METHOD

3.1. Overview

The architecture of GLE-Net is shown in Fig. 2, which follows a two-stage fusion paradigm. The first stage focuses on the feature maps generated from the image-based feature extractor (*I*) and the mmWave radar-based feature extractor (*M*), and the anchors set is generated through a RPN to achieve the complementarity of multi-modality data: $A = \{A^I, A^M\} = \{a_y\}_{y=1}^{Y}$, where $Y = |A|$ is the total number of anchors. The anchors with low confidence and large overlap are filtered to obtain the candidate set of processed anchors, $A_P = \{A_P^I, A_P^M\} = \{a_p^y\}_{y=1}^{Y}$. Meanwhile, the auxiliary module uses the generated candidate set to locally enhance the image branch. The global proposal set for both the image and mmWave radar branches can be obtained. The second stage employs the global fusion of the two sensors' features via a variant transformer. Such a two-stage fusion scheme is able to generate sufficiently robust proposals while focusing the fusion within the RoIs.

3.2. Local Enhancement

Feature extractor. GLE-Net uses two feature extractors with the same structure for camera and mmWave radar inputs. The object information of camera is relatively sufficient, and the mmWave radar point cloud is relatively sparse. For both camera and mmWave radar branches, resnet50 [6] is used as the backbone. The backbone of mmWave radar branch adopts the output of intermediate convolution layer to avoid the loss of object information due to deep convolution. The feature extractor corresponding to each branch is applied to all the corresponding inputs and generates a set of feature maps.

Anchor generator. The anchors are extracted by using a sliding window (3×3). The initial anchors are redundant and should be filtered by a two-step post-processing method. The first step is to filter out low-confidence anchors that contain

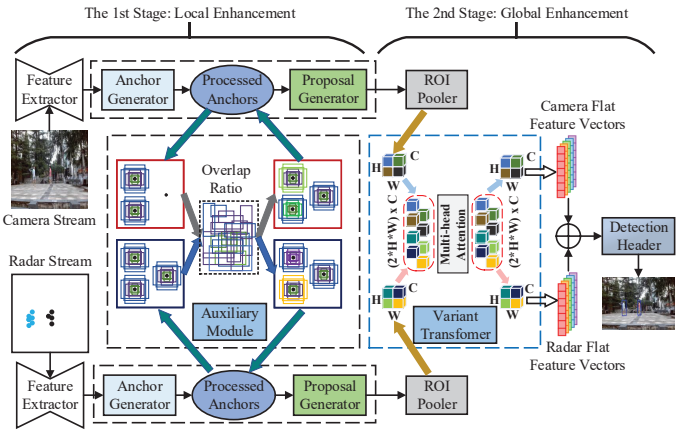


Fig. 2. Overview of GLE-Net.

any objects. The second step is to perform non-maximum suppression (NMS), which further filters out overlapping anchors. After the above operations, we can obtain the confidence ranking of all anchors, and select the top 512 to form the processed anchor set.

Auxiliary module. The camera and mmWave radar branches separately generate the processed anchors to construct an aggregated candidate set, and then the auxiliary module uses it to perform local enhancement of single branch on the anchor level. Specifically, we set the confidence threshold of the object detection boxes as 0.5. The aggregated candidate anchors are sorted in descending order according to the confidences. The anchor a_p^y with the highest confidence is selected and added to the output list. Meanwhile, a_p^y is removed from the aggregated candidate set. Candidate anchors larger than the threshold are removed by calculating the overlap ratio of a_p^y to other anchors a_p^{other}:

$$Overlap\ ratio = \frac{a_p^y \bigcap a_p^{other}}{a_p^y \bigcup a_p^{other}}. \tag{1}$$

Finally, the above process is repeated to justify the sample size. If the number of samples is less than 512, the corresponding anchors are extracted from the aggregated candidate anchors for supplementation. Otherwise, the results are selected according to the sorting of the list.

Proposal generator. After the single-branch anchor is enhanced, the generated anchors are used as the input of the proposal generator to obtain proposals for the input of the global fusion later. The proposal generator is used in *RoI poolers* to create region-wise features. For each proposal, the pooling operation is applied to the feature map of each frame to generate $2N$ $C\times W\times H$ feature tensors, where N is the number of feature maps, C is the number of channels (i.e., 512), and $W\times H$ is 2D pooling size.

3.3. Global Enhancement

The key idea of our design is to exploit the multi-head attention mechanism of the variant transformer [20] to incorporate the global information of camera and mmWave radar. The extracted feature maps adopt average pooling (1×1) in *Encoder* to obtain fine-grained input sequences since the data representation of mmWave radar is relatively simple. The variant transformer structure takes a sequence of discrete tokens as inputs, and each token is represented by a feature vector. Specifically, the variant transformer adopts linear projections to calculate a set of queries, keys and values (Q, K, and V):

$$Q = F^{in}W^q, K = F^{in}W^k, V = F^{in}W^v \tag{2}$$

where F^{in} denotes the input sequence, and W^q, W^k, W^v are the weight matrices. The attention weights of self-attention need to be calculated, which determine the degree of attention to other parts of the input vector when encoding a feature vector at a certain position. It can be obtained by doing the dot product of Q and K,

$$S = soft\max\left(\frac{QK^T}{\sqrt{D^k}}\right)V \tag{3}$$

where $\sqrt{D^k}$ is the square root of the first dimension of the weight matrix. Finally, the variant transformer exploits nonlinear transformation to calculate the output features (F^{out}):

$$F^{out} = MLP(S) + F^{in} \tag{4}$$

The output features have the same shape as the input features. The variant transformer contains multiple parallel attention "heads" (i.e., 6), which generates Q, K, and V values for each input sequence from Eq. (2) and concatenates the result values of S by Eq. (3). Note that after the variant transformer processes the input sequences, the outputs are reshaped to 7×7, and then fed to the fully connected layer to obtain the detection results.

4. EXPERIMENTS

Datasets: As shown in Fig. 3(a), we use a USB 3.0 camera and a compact commodity mmWave radar IWR6843 with 60-64 GHz [21] to collect two datasets, one with 2290 key frames (1922 for training and 368 for testing) of camera and mmWave radar data under foggy weather condition, and the other with 2112 key frames (1762 for training and 350 for testing) under poor illumination condition. During the collection, we let volunteers walk randomly in front of the collection equipment. We manually annotate each frame with 2D bounding boxes. To encompass diverse configurations, as shown in Figs. 3(b) and (c), we collect the data from 15 and 13 scenes under foggy weather and poor illumination conditions respectively, and each frame contains 1 to 4 objects. Note that we use the COCO [12] + ExDark [22] (93821 frames, 12 classes) and COCO [12] + RESIDE [23] (90782 frames, 12 classes) datasets to pre-train the image branch so that the fusion model can adapt to new scenes using less labeled multi-modality data.

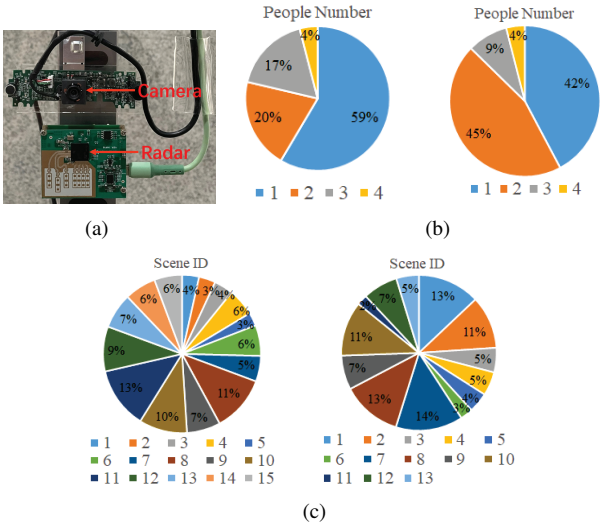


Fig. 3. (a) Our data collection equipment. (b) and (c) show the distribution of people number and collection scenes in our two datasets under foggy weather (left) and poor illumination (right) conditions.

Baselines: We compare GLE-Net with four baselines: (1) **Faster RCNN-Camera** [6], which exploits the COCO + RESIDE and COCO + ExDark datasets to pre-train faster-rcnn [5] under foggy weather and poor illumination conditions respectively, and then uses our image data to train an improved faster-rcnn [6]; (2) **Faster RCNN-Radar** [6], which uses the intermediate layer output of improved faster-rcnn [6], and is trained on our mmWave radar data; (3) **milliEye** [7], which is a fusion network based on one-stage detector; (4) **Naive Fusion** [8], which feeds camera and mmWave radar inputs into faster-rcnn [5], and then extracts proposals for fusion via the concatenation operation.

4.1. Results under Foggy Weather Condition

Baseline Comparison. As shown in Tab. 1, GLE-Net is superior to other detectors under various IoU settings. Specifically, we have three observations when IoU=0.5 (similar results with other IoU settings). *First*, Naive Fusion leverages multi-sensor data to obtain rich proposals with 1.1% and 43.2% AP improvement over detectors based on camera and mmWave radar respectively. *Second*, milliEye achieves lower AP than Naive Fusion, because milliEye does not obtain enough proposals of missing objects and only refines the detected objects. *Third*, Naive Fusion enables the detector to obtain sufficient proposals, but it ignores the filtering of redundant proposals. In contrast, GLE-Net improves the AP by 2.9% compared to Naive Fusion.

Ablation Study. We conduct ablation experiments to verify the contribution of each key module of GLE-Net. As shown in Tab. 1, GLE-Net achieves average gains of 4.5%, 4.72% and 6.5% with auxiliary module, variant transformer and both compared to Naive Fusion, respectively. The reason is that the auxiliary module can locally enhance the single-branch missing proposals in the early feature extraction stage, and the variant transformer in the late fusion stage can quantize the contribution of each branch proposal.

Methods	IoU	0.50	0.60	0.65	0.70	0.75
Faster RCNN-Camera [6]		87.5	74.3	61.9	48.2	29.8
Faster RCNN-Radar [6]		45.4	30.0	23.7	18.6	15.1
milliEye [7]		77.0	62.3	49.6	34.6	19.8
Naive Fusion [8]		88.6	83.2	78.4	70.4	62.3
GLE-Net w/o Auxiliary Module		91.2	87.4	84.2	76.8	66.9
GLE-Net w/o Variant Transformer		90.6	86.5	83.7	77.5	67.1
GLE-Net (**Ours**)		**91.5**	**88.3**	**84.9**	**80.5**	**70.2**

Table 1. Performance comparisons with four baselines and ablation experiments on the foggy weather dataset.

Methods	IoU	0.50	0.60	0.65	0.70	0.75
Faster RCNN-Camera [6]		69.6	55.5	44.8	29.4	15.8
Faster RCNN-Radar [6]		59.5	27.9	14.0	6.1	3.4
milliEye [7]		61.3	44.6	35.4	22.0	13.4
Naive Fusion [8]		70.8	56.5	49.0	36.0	22.9
GLE-Net w/o Auxiliary Module		78.4	65.2	53.6	45.4	26.1
GLE-Net w/o Variant Transformer		79.3	67.3	54.8	39.8	26.5
GLE-Net (**Ours**)		**80.5**	**67.5**	**55.2**	**46.3**	**27.4**

Table 2. Performance comparisons with four baselines and ablation experiments on the poor illumination dataset.

4.2. Results under Poor Illumination Condition

Baseline Comparison. Tab. 2 compares the competing models on the poor illumination dataset, from which we have similar observations with that on the foggy weather dataset. Specifically, Naive Fusion improves AP by 1.2% and 11.3% compared to Faster RCNN-Camera and Faster RCNN-Radar respectively. milliEye also fuses camera and mmWave radar for detection, but its AP is 9.5% lower than Naive Fusion. In contrast, GLE-Net performs local enhancement and global enhancement during the fusion process, which improves AP by 9.7% compared to Naive Fusion.

Ablation Study. As shown in Tab. 2, GLE-Net achieves average gains of 6.5%, 6.7% and 8.34% with auxiliary module, variant transformer and both compared to Naive Fusion on the poor illumination dataset, respectively.

5. CONCLUSION

We present GLE-Net to exploit complementary advantages of camera and mmWave radar via a global-local fusion for enabling robust object detection under adverse conditions. GLE-Net extracts feature maps from multi-modality data and uses an auxiliary module to process the generated anchors to locally enhance the single branch. A variant transformer is exploited to achieve global enhancement for the obtained proposals. Finally, we collect two datasets and conduct extensive experiments to verify that GLE-Net surpasses other state-of-the-art methods under both foggy weather and poor illumination conditions.

6. REFERENCES

[1] Dunja Božić-Štulić, Željko Marušić, and Sven Gotovac, "Deep learning approach in aerial imagery for supporting land search and rescue missions," *Int. J. Comput. Vision*, vol. 127, no. 9, pp. 1256–1278, 2019.

[2] Ronjie Mar L Malinao and Alexander A Hernandez, "Potentials of using unmanned aerial vehicle in the philippine farming sector: Empirical evidence from field survey," in *Proceedings of the National Conference in Information Technology Education*, 2018.

[3] Pengfei Zhu, Longyin Wen, Dawei Du, Xiao Bian, Qinghua Hu, and Haibin Ling, "Vision meets drones: Past, present and future," *arXiv preprint arXiv:2001.06303*, 2020.

[4] Tianyuan Wang, Can Ma, Haoshan Su, and Weiping Wang, "Cspn: Multi-scale cascade spatial pyramid network for object detection," in *ICASSP*, 2021, pp. 1490–1494.

[5] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Adv. neural inf. proces. syst.*, vol. 28, pp. 91–99, 2015.

[6] Masoud Jalayer, Reza Jalayer, Amin Kaboli, Carlotta Orsenigo, and Carlo Vercellis, "Automatic visual inspection of rare defects: A framework based on gp-wgan and enhanced faster r-cnn," *arXiv preprint arXiv:2105.00447*, 2021.

[7] Xian Shuai, Yulin Shen, Yi Tang, Shuyao Shi, Luping Ji, and Guoliang Xing, "millieye: A lightweight mmwave radar and camera fusion system for robust object detection," in *IoTDI*, 2021, pp. 145–157.

[8] Simon Chadwick, Will Maddern, and Paul Newman, "Distant vehicle detection using radar and vision," in *ICRA*, 2019, pp. 8311–8317.

[9] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al., "A density-based algorithm for discovering clusters in large spatial databases with noise.," in *kdd*, 1996, pp. 226–231.

[10] Bin Yang, Runsheng Guo, Ming Liang, Sergio Casas, and Raquel Urtasun, "Radarnet: Exploiting radar for robust perception of dynamic objects," in *ECCV*, 2020, pp. 496–512.

[11] Meet Shah, Zhiling Huang, Ankit Laddha, Matthew Langford, Blake Barber, Sidney Zhang, Carlos Vallespi-Gonzalez, and Raquel Urtasun, "Liranet: End-to-end trajectory prediction using spatio-temporal radar fusion," *arXiv preprint arXiv:2010.00731*, 2020.

[12] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick, "Microsoft coco: Common objects in context," in *ECCV*, 2014, pp. 740–755.

[13] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman, "The pascal visual object classes challenge: A retrospective," *Int. J. Comput. Vision*, vol. 111, no. 1, pp. 98–136, 2015.

[14] Mario Bijelic, Tobias Gruber, Fahim Mannan, Florian Kraus, Werner Ritter, Klaus Dietmayer, and Felix Heide, "Seeing through fog without seeing fog: Deep multimodal sensor fusion in unseen adverse weather," in *CVPR*, 2020, pp. 11682–11692.

[15] Danfei Xu, Dragomir Anguelov, and Ashesh Jain, "Pointfusion: Deep sensor fusion for 3d bounding box estimation," in *CVPR*, 2018, pp. 244–253.

[16] Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li, "Pointrcnn: 3d object proposal generation and detection from point cloud," in *CVPR*, 2019, pp. 770–779.

[17] Kun Qian, Shilin Zhu, Xinyu Zhang, and Li Erran Li, "Robust multimodal vehicle detection in foggy weather using complementary lidar and radar signals," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 444–453.

[18] Ramin Nabati and Hairong Qi, "Rrpn: Radar region proposal network for object detection in autonomous vehicles," in *ICIP*, 2019, pp. 3093–3097.

[19] Ramin Nabati and Hairong Qi, "Radar-camera sensor fusion for joint object detection and distance estimation in autonomous vehicles," *arXiv preprint arXiv:2009.08428*, 2020.

[20] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Adv. neural inf. proces. syst.*, 2017, pp. 5998–6008.

[21] Zhen Meng, Song Fu, Jie Yan, Hongyuan Liang, Anfu Zhou, Shilin Zhu, Huadong Ma, Jianhua Liu, and Ning Yang, "Gait recognition for co-existing multiple people using millimeter wave sensing," in *AAAI*, 2020, pp. 849–856.

[22] Yuen Peng Loh and Chee Seng Chan, "Getting to know low-light images with the exclusively dark dataset," *Comput. Vis. Image Und.*, vol. 178, pp. 30–42, 2019.

[23] Boyi Li, Wenqi Ren, Dengpan Fu, Dacheng Tao, Dan Feng, Wenjun Zeng, and Zhangyang Wang, "Benchmarking single-image dehazing and beyond," *IEEE Trans. Image Process*, vol. 28, no. 1, pp. 492–505, 2018.