Poster number: 3433

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,

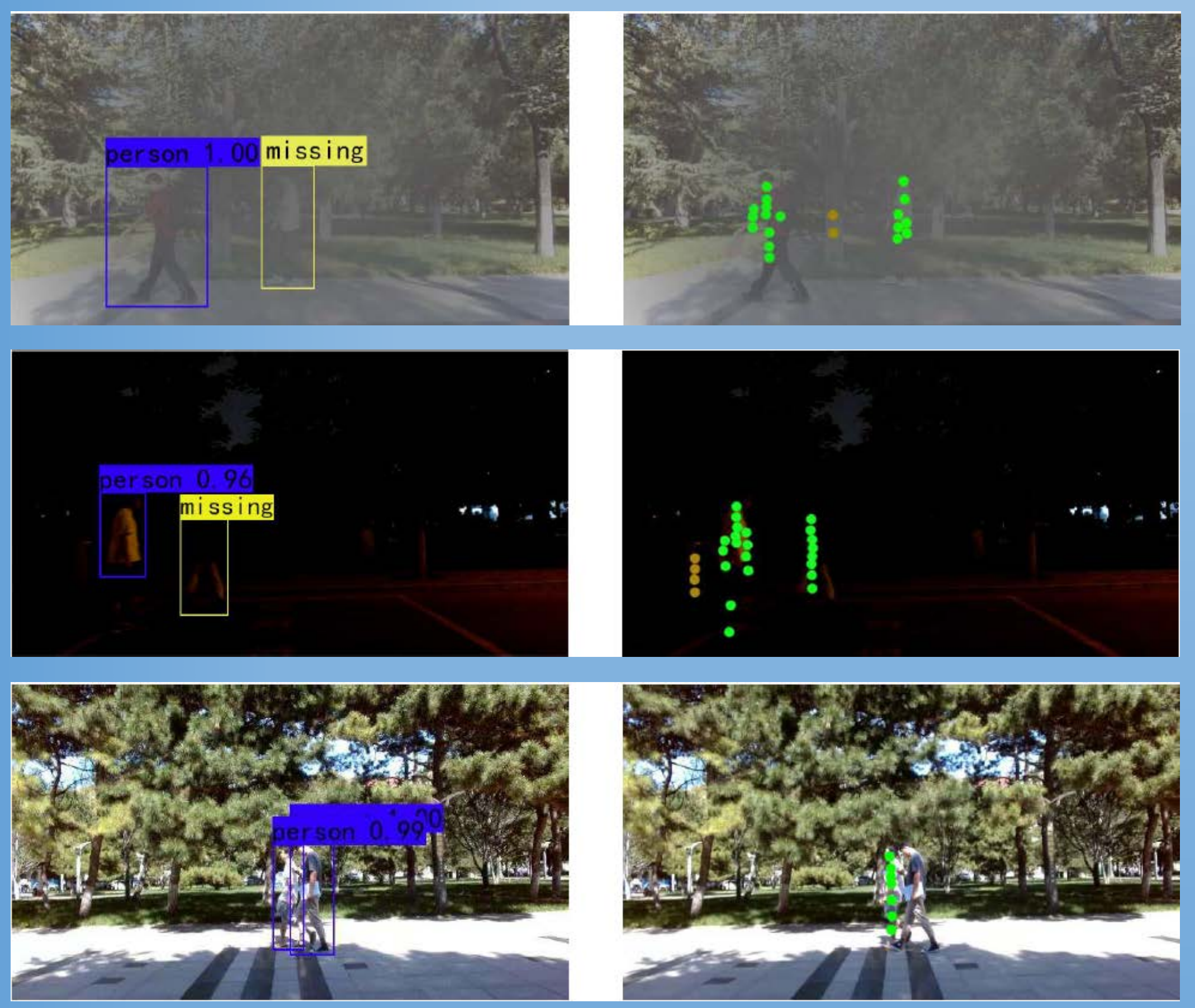
Beijing University of Posts and Telecommunications, Beijing 100876, China

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).

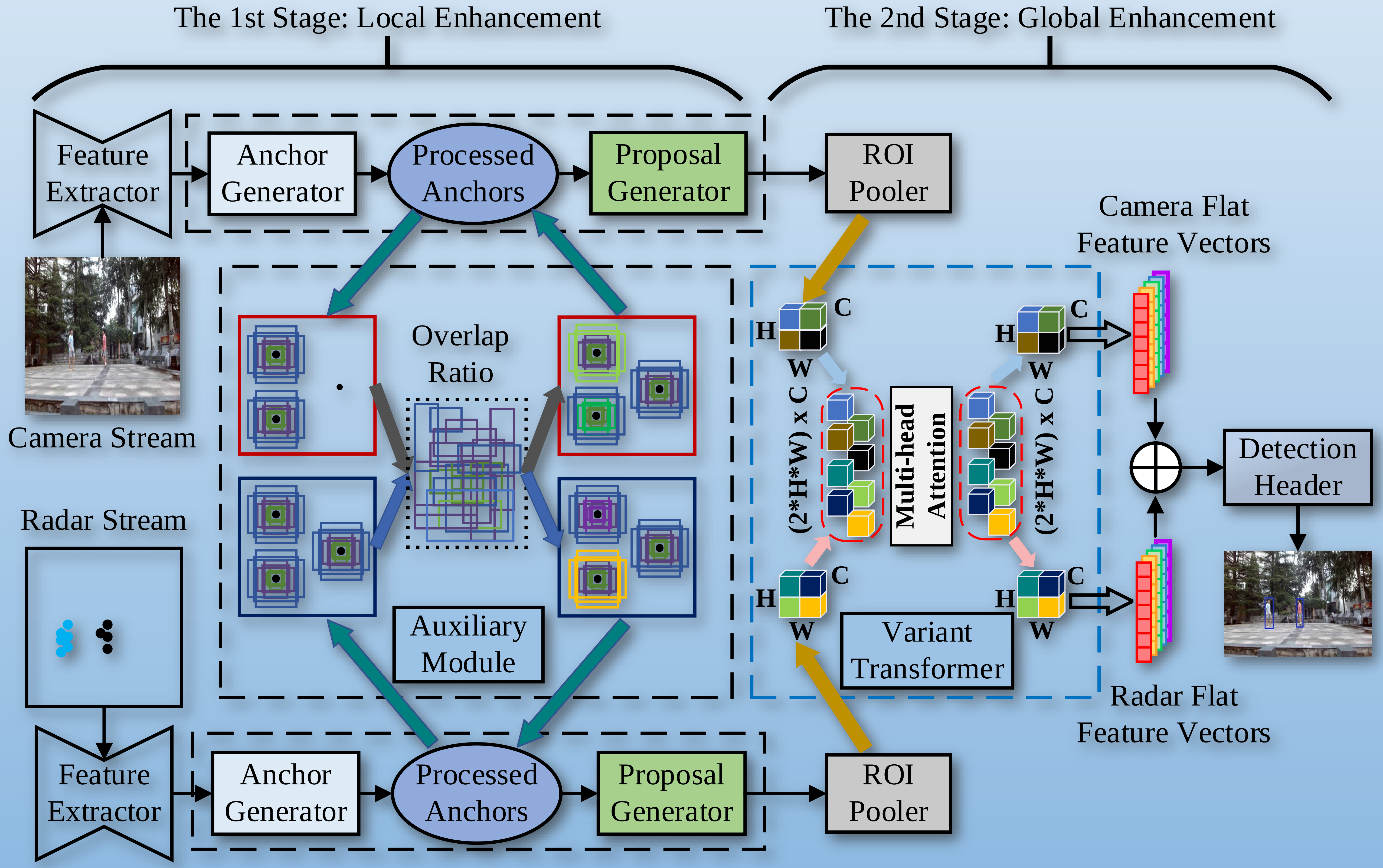
Introduction

- The deep learning has promoted the application of object detection in many fields, such as search and rescue, farmland thief detection, and traffic monitoring.



- 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.
- 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.

Method (GLE-Net)



- **Feature extractor**. Resnet50 is used as the backbone. The backbone of mmWave radar branch adopts the output of intermediate convolution layer.
- **Anchor generator**. The anchors are extracted by using a sliding window (3×3). We select the top 512 to form the processed anchor set.
- **Auxiliary module**. The aggregated candidate anchors are sorted in descending order according to the confidences. Candidate anchors larger than the threshold are removed by calculating the overlap ratio of a_p^y to other anchors a_p^{other}:

$$Overlap\ Ratio = \frac{a_p^y \cap a_p^{other}}{a_p^y \cup a_p^{other}}$$

- **Proposal generator**. The generated anchors are used as the input of the proposal generator to obtain proposals for the input of the global fusion later.
- **RoI poolers**. For each proposal, the pooling operation is applied to the feature map of each frame to generate feature tensors.
- **Variant Transformer.**
  - ✓ *First*, the extracted feature maps adopt average pooling (1×1) in Encoder to obtain fine-grained input sequences. Specifically, the variant transformer adopts linear projections to calculate a set of queries, keys and values (*Q*, *K*, and *V*).

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

  - ✓ *Second*, 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.

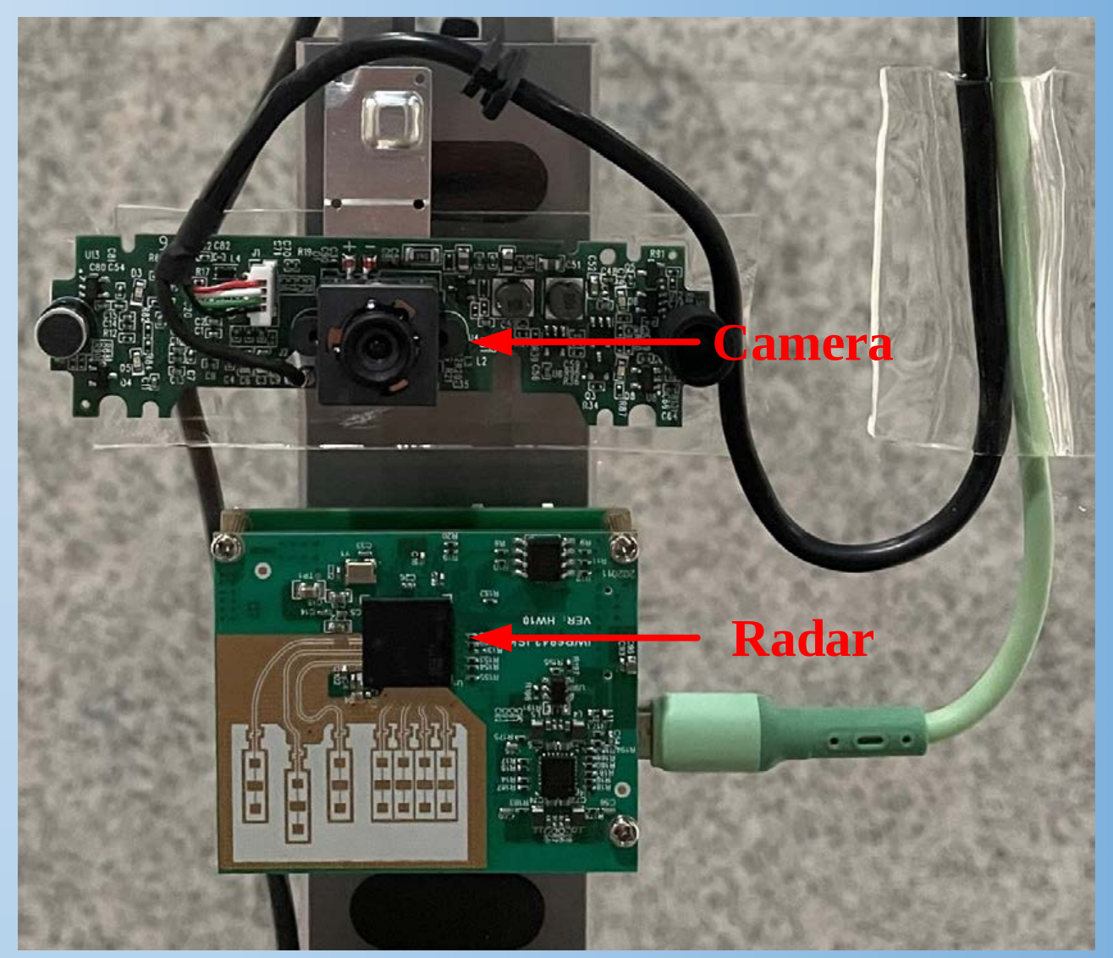
$$S = softmax(\frac{QK^T}{\sqrt{D^k}})V$$

  - ✓ *Third*, the variant transformer exploits nonlinear transformation to calculate the output features (F^{out}).

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

Experiments

- Data collection equipment: A USB 3.0 camera and a compact commodity mmWave radar IWR6843



Methods	IoU	0.50	0.60	0.65	0.7	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]		<u>88.6</u>	<u>83.2</u>	<u>78.4</u>	<u>70.4</u>	<u>62.3</u>
GLE-Net w/o Auxiliary Module		<u>91.2</u>	<u>87.4</u>	<u>84.2</u>	76.8	66.9
GLE-Net w/o Variant Transformer		90.6	86.5	83.7	<u>77.5</u>	<u>67.1</u>
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	0.7	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]		<u>70.8</u>	<u>56.5</u>	<u>49.0</u>	<u>36.0</u>	<u>22.9</u>
GLE-Net w/o Auxiliary Module		78.4	65.2	53.6	<u>45.4</u>	26.1
GLE-Net w/o Variant Transformer		<u>79.3</u>	<u>67.3</u>	<u>54.8</u>	39.8	<u>26.5</u>
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.

Conclusion

- 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
- We collect two datasets and verify GLE-Net. 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 and Naive Fusion, respectively, under foggy weather condition, and improves by 19.2% and 9.7% respectively under poor illumination condition.