2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

Automatic Captioning based on Visible and Infrared Images

Yan Wang, Shuli Lou, Kai Wang, Yunzhe Wang, Xiaohu Yuan and Huaping Liu*

***Abstract*— In this paper, we tackle the task of image captioning with the complementarity of visible light images and infrared images. To address this problem, we propose an RGB-IR image fusion captioning model, which can take full advantage of visible light images and infrared images under different conditions. Meanwhile, we develop a wearable environment-assisted system. In addition, we collect and annotate a new dataset containing 3510 pairs of RGB-IR images to support model training. Finally, we conduct extensive experiments to evaluate the model and system. Experimental results show that our new method and system significantly outperform baselines on multiple metrics and have potential practical value.**

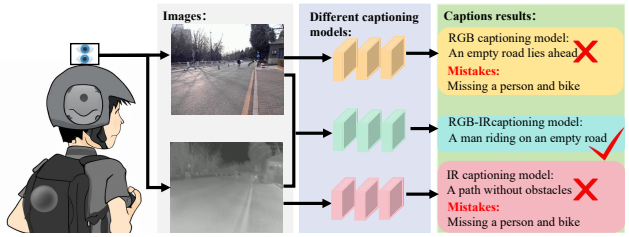

Fig. 1. In a real environment, a visible light image and an infrared image are acquired through dual cameras. With the benefits of RGB-IR fusion, better captioning results can be obtained.

I. INTRODUCTION

Image captioning aims to translate visual features of images into natural language descriptions, enabling computers to generate captions similar to humans. With the rapid development in this field, image captioning can not only help computers understand image content, but also be applied to scenarios such as assisting visually impaired individuals in accessing image information.

Commonly used methods for image captioning combine convolutional neural networks (CNNs) and recurrent neural networks (RNNs) into encoder-decoder frameworks, see [1], [2] for related research. Subsequent studies explore and improve this framework by introducing attention mechanism-based models, such as [3], [4], [5], [6]. Ref.[7] Optimizes the structure of the generated captions. Ref.[8] Not only enriches the generated captions, but also makes them more distinguishable. Some studies also employ reinforcement learning to generate captions more aligned with human approval [9], [10], [11]. Ref.[12] Makes the generated captions closer to conforming to human language structures. Ref.[13] proposes a framework that incrementally updates the scene graph using object detection and connects captions from different frames to generate more accurate and thorough image captions. Experiments on AI2THOR dataset and real robots validate its effectiveness. In military decision-making and tactical planning, Ref.[14] emphasizes the importance of image features and information from captions to assist subsequent human actions. Ref.[15] proposes MLCA-Net with multi-level and contextual attention for more flexible and diverse caption generation. Ref.[16] proposes a structured-attention method for high-resolution remote sensing image captioning with higher accuracy and ability to generate semantic content and segmentation masks. Ref.[17] proposes a multi-scale feature fusion method with denoising for remote sensing image captioning, helping obtain denoised multi-scale features in encoder-decoder frameworks.

Currently, image captioning tasks focus on accurately extracting features from visible light images, but insufficient light or light pollution limits their performance. In contrast, infrared images perform better in these environments due to their sensitivity to thermal targets. Ref.[18] proposes an infrared image captioning model, which demonstrates this. However, infrared images may not match the quality of visible light images in well-lit environments. In existing work, visible light images and infrared images have not been combined together for use in image captioning tasks.

To enhance the generalizability of image captioning applications to diverse real-world scenarios, we propose an innovative RGB-IR image captioning model that fully utilizes the complementarity between visible light images and infrared images, unlike traditional single-modal image captioning models, our model leverages both visible light and infrared images to generate more accurate image captions. In addition, we construct a multimodal dataset for model training. Finally, following the design ideas for wearable devices from [18], [19], [20], [21], we design a wearable device and deploy visible light sensors, infrared sensors, and an RGB-IR image captioning model, achieving real-time RGB-IR image caption generation as shown in Fig. 1. Our wearable device can perform real-time environmental perception tasks, not only providing assistance for the visually impaired, but also useful for outdoor exploration and rescue operations. Overall, the main contributions of this paper are as follows:

- We develop a new RGB-IR image captioning model that can effectively utilize the complementary information in visible light images and infrared images to generate more informative and precise image captions.

Yan Wang, Shuli Lou and Kai Wang are with Department of Physics and Electronic Information, Yantai University. Yunzhe Wang is with La Jolla County day school. Xiaohu Yuan and Huaping Liu are with Department of Computer Science and Technology, Tsinghua University. This work was completed while Yan Wang, Kai Wang and Yunzhe Wang were visiting Tsinghua University. This work was supported in part by the National Natural Science Fund under Grant 62025304.

*Corresponding Author: Huaping Liu(hpliu@tsinghua.edu.cn).

- We creat a multimodal dataset containing 3510 pairs of visible and infrared images for training the image captioning model. This dataset provides a rich data resource that helps to improve the accuracy and generalisation of the model.
- We develop a wearable device helmet and deploy a trained model to the helmet to enable real-time environmental perception. This innovation enables users to obtain descriptive information about their surroundings, enhancing their perception and safety on the move. Furthermore, we perform comprehensive experiments in real-world scenarios to validate the efficacy of our proposed approach and model.

In Section II, we review related work on infrared image captioning and multimodal image captioning. Section III outlines the problems faced in our work. The system utilized is described in Section IV. The methods for dataset collection and annotation are explained in Section V. The principles underpinning our proposed RGB-IR image captioning model are elaborated on in Section VI. Section VII details the experimental design and analyzes the results obtained. The physical verification results are presented in Section VIII. Finally, Section IX concludes the paper.

II. RELATED WORK

A. Infrared Image Captioning

Traditional images still face difficulties in understanding complex scenes and targets. Infrared imaging makes infrared images an effective way to obtain target radiation information. Infrared image captioning is receiving increasing attention from researchers. Our previous work [18] uses CycleGAN to convert visible images to infrared and designs a non-infrared feature filter to construct the infrared image caption dataset IR-MSCOCO. On this basis, the image captioning framework in [8] is adopted for validation. The trained model is deployed on wearable devices for real environment evaluation. Ref.[22] performs image captioning on infrared image datasets using YOLOv6 and LSTM encoding-decoding, and optimizes domain migration to improve detector adaptability. The key is an object-directed attention mechanism that combines detector semantic information and image features, and weights the features proportionally when generating words in decoding to produce descriptions highly relevant to predefined objects.

B. Fused Image Captioning

With the increase of multi-source heterogeneous data, effectively fusing different modal information has become an important research direction for image captioning tasks. Early image captioning models are mainly based on single visual information and had limited semantic expression capabilities. To enhance models' image understanding and description capabilities, many subsequent works have explored multimodal fusion. For example, Ref.[23] uses a multimodal fusion method for video description tasks. In addition to commonly used image features and motion features, it additionally introduces audio features to provide extra content

Fig. 2. The developed helmet for automatic captioning.

cues to help generate richer descriptions. Ref.[24] proposes a deep learning-based ultrasound image captioning method using object detection. This method fuses an object detection module in the encoding stage. It can not only locate the positions of focal areas in ultrasound images but also carry less noise information and capture more detailed information about the focal areas in the encoding vectors of focal areas. This work explores the effectiveness of object detection frameworks in medical image description tasks.

III. DESCRIPTION OF THE PROBLEM

The task of this paper is to acquire visible light images and infrared images simultaneously in real-world environments and input them into an image captioning model to describe the current environment. However, in the field of image captioning, there is a lack of visible-infrared datasets, and existing datasets are insufficient to support research on multimodal image captioning tasks.

In the process of building a multimodal dataset, visible light images and infrared images need to be collected across multiple scenes. Since the two types of images have different resolutions, they cannot be directly used for multimodal image fusion captioning tasks, and alignment work needs to be done on the dataset.

After obtaining visible light images and infrared images of the same scene, suitable fusion methods need to be developed to fuse the visible light images and infrared images, in order to maximize the utilization of information provided by each of these modalities, to accomplish the task of environmental recognition and description.

IV. PROTOTYPE SYSTEM DEVELOPMENT

To test the feasibility of our approach in a multimodal image scene task, we design a multimodal data acquisition system that combines a visible camera with an infrared camera. We then use this system for real-time image captioning tests. In this section, we explain our process in detail.

A. Hardware

Our prototyping system consists of an ergonomic helmet, a 28,000mAh portable power supply, an IR-Pilot640 infrared camera, an SG1-AR0144C-8310-GMSL visible camera, an NVIDIA Jetson TX2 processor, Sony headphones, a mini Bluetooth controller, and a backpack, as shown in Fig. 2. Different from our previous work [18], which only contains the infrared camera, we place the processor in the backpack

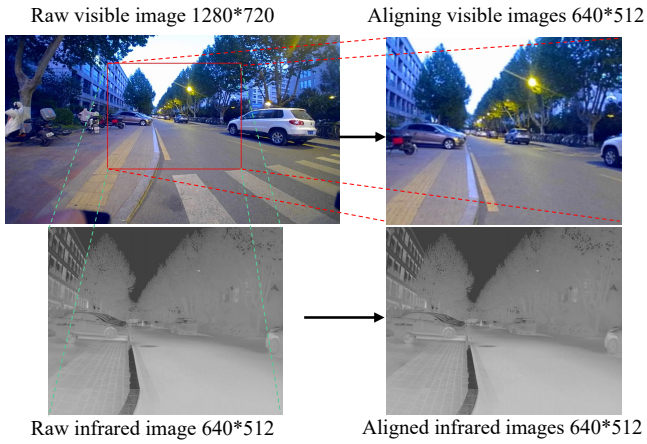

Fig. 3. Example before and after image alignment.

instead of on the helmet to reduce the weight of the helmet, and mount the infrared and visible light cameras at the top of the helmet. Our model is deployed on the processor. The user controls the infrared and visible light cameras to capture current environmental scene maps through the Bluetooth controller. The results are processed by the processor and output to the headphones through the integrated image-to-speech conversion module.

B. Aligment

The visible light camera we use has a resolution of 1280×720, while the infrared camera has a resolution of 640×512. In order to align the visible-infrared frames, we refer to the method of homography of the ground plane[25]. First, based on the four corner points of the infrared image, the four corresponding corner points of the same scene are marked in the visible light image. Then, the image content between these four corner points in the visible light image is cropped, finally resulting in a visible light image with the same scene as the infrared image and a resolution of 640×512 pixels, as shown in Fig. 3. Compared to geometric calibration of cameras, the method we use is more convenient, avoiding the cumbersome calibration process, camera parameter estimation and determination of relative position relationships, reducing the complexity and difficulty of system deployment, especially in real-world scenarios where devices need to be frequently adjusted and moved. Although there may be slight alignment errors during image fusion, such errors are acceptable and tolerable for the application field of image caption generation. Rather than pursuing absolute precise positional alignment, we care more about the semantic alignment between visible and infrared cameras.

V. Dataset Collection and Annotations

We use the developed multimodal data acquisition system to capture outdoor videos and extract images by frame in these videos. In this section, we explain our dataset creation process in detail.

Fig. 4. Example Dataset.

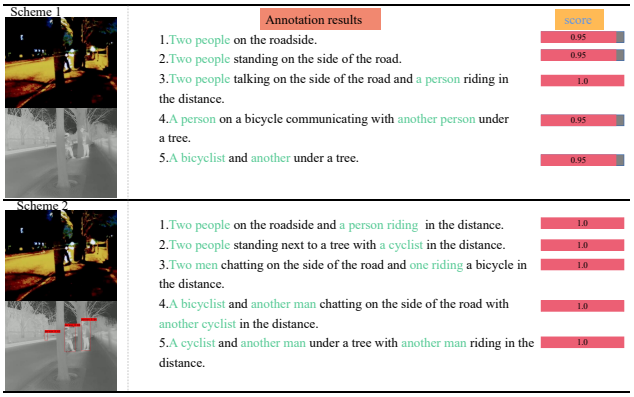

Fig. 5. Annotations programme and scores.

A. Dataset Collection

We invite volunteers to wear the multimodal data collection device. In this device, the visible and infrared cameras capture video with a resolution of 640 × 512 and fps of 30 frames. We collect data from 13 outdoor scenes. These scenes are categorized into three high-level categories - daytime, nightfall and night - and four scenarios (campus, city street, country path, and field) are distributed among these two categories. The average duration of the videos is 3 minutes, with a total frame count of 140,400 frames, of which 70,200 frames each are visible image frames and infrared image frames. To ensure the diversity of images, we save one image every 20 frames for visible and infrared videos, respectively, and obtain a total of 3,510 visible images and 3,510 infrared images, hereinafter referred to as image pairs, to form a multimodal dataset. It is divided into training set, validation set and test set in the ratio of 8:1:1. Fig. 4 shows an example of the visualization of the dataset.

B. Annotations

Our goal is to annotate 5 captions for each image pair. Section V-B.1 introduces the annotation requirements and schemes for the dataset; Section V-B.2 introduces the evaluation of the schemes.

1) Annotations Programme: We first attempt to fuse the visible light images and infrared images at the pixel level through algorithms to generate new fused images, which are

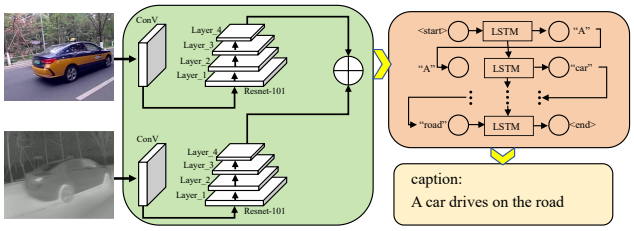

Fig. 6. The proposed RGB-IR image captioning framework structure. It specifically shows the working details of fusing the visible light image and infrared image. The fused feature maps are decoded by the decoder to output image captions.

then manually annotated. However, this method requires very high image alignment quality. Our tests find that the resulting images have poor effects, with blurred pixel information and unclear object boundaries, posing great difficulties for subsequent annotation work. We finally decide to abandon this direct pixel fusion approach. We then come up with two other feasible solutions. Scheme 1 refers to [26] and directly annotates the image pairs. Since the multimodal dataset contains daytime, nightfall and night scenes, when annotating night scene images, we mainly rely on infrared images. However, infrared images have fewer texture details and targets are not easy to locate. Therefore, we propose Scheme 2, which first performs target detection on infrared images to obtain target locations, and then the labeled infrared images and visible light images are annotated. Our two annotation schemes based on [18], [26] both have the following requirements:

- *Do not use colour words.*
- *Do not start with 'There is'.*
- *Do not use temporal words.*
- *Do not describe what a person may say.*
- *Do not describe things that may happen in the future or past.*

2) Validation Programme: To verify the effectiveness of Scheme II, we design a validation experiment. We randomly select 50 pairs of images from the entire dataset and hire five annotators to annotate according to scheme I and another five annotators to annotate according to scheme II. Eventually, we scored the captions labelled by the two groups of annotators to evaluate the advantages and disadvantages of the two schemes. In addition, we use statistical hypothesis testing and calculate t-values and p-values to determine the significance of the differences and to rule out chance.

The final results of both schemes are 50 pairs of images, totalling 250 captions. We score the 250 captions from each scheme separately. We score based on the number of missed targets, taking people, cars, and bicycles as targets, deducting 0.05 points for each missed target. A scoring example is shown in Fig. 5. We finally obtain a p-value of 1.75×10^{-6} and a t-value of 5.42, with Scheme 2 having a higher average score than Scheme 1. Therefore, We annotate the multimodal dataset according to Scheme 2, annotating 5 captions for each image pair, finally resulting in a total of 17,550 captions.

VI. Methods

The RGB-IR image captioning framework architecture is shown in Fig. 6. Our model contains two main components: an image fusion network and a captioning generation model, which are used for multimodal image feature fusion and mapping features to captions, respectively. In the training part, we adopt the method and parameters suggested in [8] for the captioning generation model.

A. Image Fusion Network

Currently, mainstream visible image caption generators are susceptible to ambient lighting and complex backgrounds. Infrared (IR) images can make up for the defects of visible light images due to their insensitivity to lighting conditions, while a single IR image has poor texture information. To address the above issues, following the methods and ideas in our previous work [27], we design an image fusion structure, enabling the image caption generator to have the advantages of both visible light image and infrared image caption generators simultaneously.

We use an image feature level fusion strategy. First, the visible light image and infrared image are input into two identical Conv+BN+ReLU and Resnet-101 networks respectively. Although the visible light image is a 3-channel RGB image and the infrared image is a single-channel one, after feature extraction through the convolutional layers, both will obtain multi-channel feature maps, so the number of channels can match. After convolution and activation functions, feature maps are extracted. Then, fusion is performed at the highly abstract feature layer level of Resnet-101. Specifically, the features output by Resnet-101 for the visible light image and infrared image are summed at the same spatial location, to obtain the fused feature map.

B. Caption Generation Model

Since the image captioning model proposed in [8] is highly modular, can be applied to different retrieval models or different caption generators, and is able to improve the discriminability of the captions, we use this image captioning model. We define M as a pair of visible light and infrared images (V,I). Then, the conditional probability model for generating a caption c given image pair M can be represented as:

$$p(c|M;\theta) = \prod_t p(w_t|w_0,\cdots,w_{t-1},M;\theta) \tag{1}$$

where $c = (w_0, w_1, \ldots, w_T)$, described in more detail in [9]. w_t is the next generated word, which depends on the previous $t-1$ words and the image feature y, and θ denotes the model parameters.

We then take the fused image feature y of the image pair M and the caption feature $g(c)$ of the caption c, and train an image pair-caption matching model that maps the image pair and caption into a common semantic space and computes a matching score $s(M,c)$ based on the cosine similarity between the two vectors.

$$s(M,c) = \frac{y \cdot g(c)}{||y||||g(c)||} \tag{2}$$

Finally, a discriminative loss function is used:

$$L_{CON}(c,M) = \max_{c'}[\alpha + s(M,c') - s(M,c)]_+ + \max_{M'}[\alpha + s(M',c) - s(M,c)]_+ \quad (3)$$

where $[x]_+ \equiv max(x,0)$. (M,c) denotes a correctly matched image-caption pair, and (M',c) and (M,c') denote incorrectly matched pairs (e.g., c' is a caption that does not describe image M). This loss function aims to make the matching score α of the correct match as high as possible above the scores of incorrect matches.

VII. EXPERIMENTS AND RESULTS

The main purpose of the experiment is to assess the performance and efficacy of image captioning models. In the test set portion of the dataset, we evaluate three different image captioning models, compare the quantitative and qualitative results of these three models, and analyze their advantages and disadvantages. We introduce the experimental design in Section VII-A and present the experimental results in Section VII-B.

A. Experimental Design

1) Score: We employ a set of standard quantitative metrics to evaluate image captions, which draw on measurement methods from machine translation, including BLEU [28], ROUGE [29], and CIDEr [30]. These three metrics each have different focuses. BLEU emphasizes precise n-gram matching, ROUGE considers word order information, while CIDEr evaluates how well the caption covers the semantic aspects of the image. Using the three evaluation metrics together can assess the quality of image captions more comprehensively.

2) Model Comparison Experimental Design: We obtain the RGB-IR captioning model on the test set of the multimodal dataset. In addition, we test the RGB captioning model using visible light images in the dataset, which is based on the image captioning model in the literature [8]. Also, we test the IR captioning model using infrared images in the dataset, which is adopted from the image captioning model in our previous work [18].

B. Experimental Results

1) Model Comparison Experimental Results: We report the scores of the models on the test set in Table I. By comparing the performance of the three models on image captioning tasks, we can see that the fusion of multimodal information has a significant effect on improving the quality of image caption generation. The RGB-IR captioning model performs markedly better than the single modal RGB captioning model and IR captioning model in terms of vocabulary matching, phrase matching of different lengths, and semantic consistency. Specifically, in terms of vocabulary matching, the RGB-IR captioning model consistently outperforms the other two models on BLEU scores from 1-gram to 4-gram, which shows that the model fused with RGB and IR visual information can generate captions that are more adherent to the reference sentences at the lexical level. In terms of phrase matching, the RGB-IR captioning model also has higher ROUGE scores than the other models, indicating that the generated image captions contain more varying length phrases consistent with the reference sentences. Most importantly, on the CIDEr metric that evaluates semantic consistency, the RGB-IR captioning model also outperforms the other models, which verifies that after fusing multimodal information, the model understands the image content more comprehensively and accurately and can generate more semantically accurate captions.

TABLE I

PERFORMANCE SCORES FOR EACH MODEL

	Methods					
Image eval type	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE	CIDEr
RGB Captioning Model for RGB Image	0.862	0.731	0.618	0.518	0.631	0.703
IR Captioning Model for IR Image	0.865	0.732	0.614	0.510	0.628	0.698
RGB-IR Captioning Model for Image	0.897	0.795	0.697	0.607	0.657	0.835

Fig. 7. Image Caption Result.

Fig. 7 shows some representative images from our test set experiments and their corresponding generated captions. It can be clearly seen that for various outdoor scenes, whether during the day or at night, the RGB-IR image captioning model is able to generate semantically accurate captions that match the scene. This is mainly thanks to the model's ability to achieve comprehensive perception and precise understanding of various scenes by fusing RGB images and IR images. The RGB images provide rich visual information such as colors, textures, shapes, etc., while the IR images can capture thermal targets and scene outlines

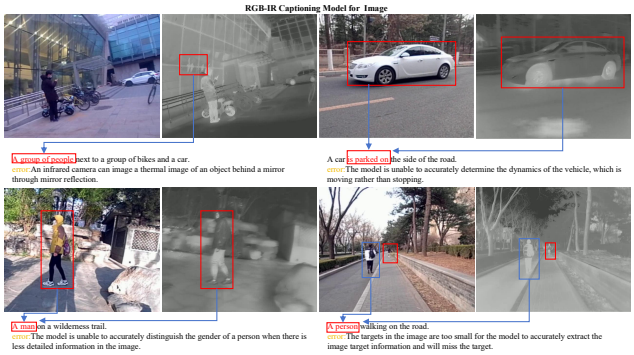

Fig. 8. Some faliure cases.

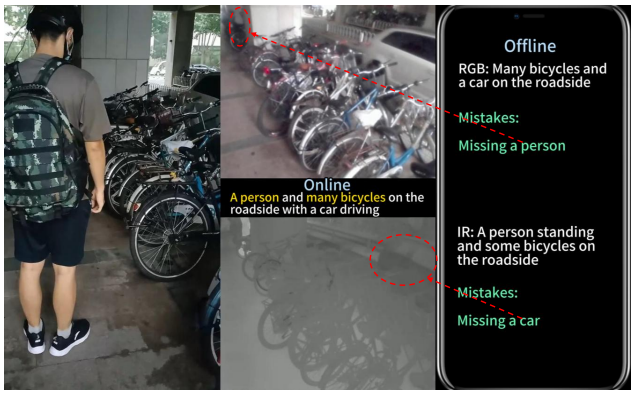

Fig. 9. In real-world testing, after pressing the button, the fused caption (with audio) appears in the middle, which is the result generated online by the RGB-IR captioning model, while the right column shows our offline processed results, containing two types of results generated by the RGB captioning model and infrared captioning model, and error analysis has been performed.

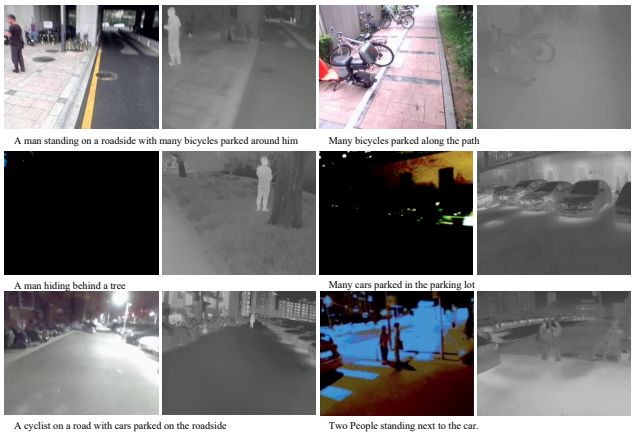

Fig. 10. Real-World Testing Results.

under nighttime and insufficient lighting conditions. The two complement each other. In contrast, conventional captioning models that rely solely on a single modality are prone to be affected by lighting changes, and have obvious defects in the semantic correctness and robustness of their generated captions, making it difficult for them to meet the needs of complex and diverse real-world applications.

2) Bias and Analysis: In the RGB-IR image captioning task, we encounter some inaccurate or inconsistent examples, as shown in Fig. 8. We describe these errors in detail and conduct a preliminary analysis on their causes. Specifically, when there are smooth reflective surfaces like mirrors in front of the infrared camera, the mirror reflects the image of the target, resulting in additional detected targets in the infrared image and thus leading to more targets in the final count than actual. Also, for vehicle targets, the system sometimes fails to distinguish whether the vehicle is stationary or moving, since it is difficult to determine the motion status of vehicles from a single static image. The human eye faces the same difficulty, needing to observe continuous image sequences to judge vehicle behavior. In addition, when the features of people are not salient in the image, the model fails to capture sufficient discriminative human body features to infer gender, which also reduces the accuracy of captions. Finally, the model does not pay enough attention to overly tiny targets, which directly leads to the failure in detecting those targets, and thus missing them in the generated captions. We believe that correlating target and environmental information, as well as incorporating object detection networks for extracting small objects, can further improve the accuracy of image captions.

VIII. Physical Verification

Recall that our motivation for introducing this objective is to address people's environmental perception issues in outdoor scenes through image captioning. Therefore, our focus is on validating the model's practicality in real-world environments, we invite multiple volunteers to wear the helmet device, and integrate the well-trained RGB-IR image captioning model into it, then test it in various outdoor scenes, as shown in Fig. 9. Fig. 10 shows some image samples from real scenes and their corresponding captions. In this way, our model goes from laboratory settings to the real world, effectively validating the practicality of our designed model and device in real environments.

The helmet device worn by volunteers is the device described in Section IV-A. After deploying the image captioning model on it and through testing, the average time from when the volunteers press the button to when they hear the sound is 1.75 seconds.

IX. Conclusions

In this paper, we develop an RGB-IR image captioning model to achieve all-weather scene understanding by utilizing the advantages of both modalities. We also construct a multimodal dataset which can support RGB-IR image captioning research. This dataset can also be applied to assistive systems for the blind to help them better perceive surrounding pedestrians and improve action safety. Moreover, the dataset can be used in urban security, outdoor exploration and other fields. Finally, we train the model using multimodal datasets and integrate it into wearable devices to perform real-time tasks in outdoor scenes. Comprehensive real-world tests demonstrate the efficacy of our model and wearable devices.

REFERENCES

[1] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3156–3164, 2015.

[2] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3128–3137, 2015.

[3] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6077–6086, 2018.

[4] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *International conference on machine learning*, pp. 2048–2057, PMLR, 2015.

[5] J. Lu, C. Xiong, D. Parikh, and R. Socher, "Knowing when to look: Adaptive attention via a visual sentinel for image captioning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 375–383, 2017.

[6] M. A. Al-Malla, A. Jafar, and N. Ghneim, "Image captioning model using attention and object features to mimic human image understanding," *Journal of Big Data*, vol. 9, no. 1, pp. 1–16, 2022.

[7] J. Lu, J. Yang, D. Batra, and D. Parikh, "Neural baby talk," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7219–7228, 2018.

[8] R. Luo, B. Price, S. Cohen, and G. Shakhnarovich, "Discriminability objective for training descriptive captions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6964–6974, 2018.

[9] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, "Self-critical sequence training for image captioning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7008–7024, 2017.

[10] U. Honda, T. Watanabe, and Y. Matsumoto, "Switching to discriminative image captioning by relieving a bottleneck of reinforcement learning," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1124–1134, 2023.

[11] P. Devi, V. Thrivikraman, D. Kashyap, and S. Shylaja, "Image captioning using reinforcement learning with bluder optimization," *Pattern Recognition and Image Analysis*, vol. 30, pp. 607–613, 2020.

[12] S. Chen, Q. Jin, P. Wang, and Q. Wu, "Say as you wish: Fine-grained control of image caption generation with abstract scene graphs," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9962–9971, 2020.

[13] X. Li, D. Guo, H. Liu, and F. Sun, "Robotic indoor scene captioning from streaming video," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6109–6115, IEEE, 2021.

[14] D. Ghataoura and S. Ogbonnaya, "Application of image captioning and retrieval to support military decision making," in *2021 international conference on military communication and information systems (ICMCIS)*, pp. 1–8, IEEE, 2021.

[15] Q. Cheng, H. Huang, Y. Xu, Y. Zhou, H. Li, and Z. Wang, "Nwpu-captions dataset and mlca-net for remote sensing image captioning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–19, 2022.

[16] R. Zhao, Z. Shi, and Z. Zou, "High-resolution remote sensing image captioning based on structured attention," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–14, 2021.

[17] W. Huang, Q. Wang, and X. Li, "Denoising-based multiscale feature fusion for remote sensing image captioning," *IEEE Geoscience and Remote Sensing Letters*, vol. 18, no. 3, pp. 436–440, 2020.

[18] C. Gao, Y. Dong, X. Yuan, Y. Han, and H. Liu, "Infrared image captioning with wearable device," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 8187–8193, IEEE, 2023.

[19] A. Olwal, K. Balke, D. Votintcev, T. Starner, P. Conn, B. Chinh, and B. Corda, "Wearable subtitles: Augmenting spoken communication with lightweight eyewear for all-day captioning," in *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*, pp. 1108–1120, 2020.

[20] A. Padmanabha, Q. Wang, D. Han, J. Diyora, K. Kacker, H. Khalid, L.-J. Chen, C. Majidi, and Z. Erickson, "Hat: Head-worn assistive teleoperation of mobile manipulators," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 12542–12548, IEEE, 2023.

[21] F. Digiacomo, A. S. Afroz, R. Pelliccia, F. Inglese, M. Milazzo, and C. Stefanini, "Head-mounted standalone real-time tracking system for moving light-emitting targets fusing vision and inertial sensors," *IEEE Transactions on Instrumentation and Measurement*, vol. 69, no. 11, pp. 8953–8961, 2020.

[22] J. Lv, T. Hui, Y. Zhi, and Y. Xu, "Infrared image caption based on object-oriented attention," *Entropy*, vol. 25, no. 5, p. 826, 2023.

[23] C. Hori, T. Hori, T.-Y. Lee, Z. Zhang, B. Harsham, J. R. Hershey, T. K. Marks, and K. Sumi, "Attention-based multimodal fusion for video description," in *Proceedings of the IEEE international conference on computer vision*, pp. 4193–4202, 2017.

[24] X. Zeng, L. Wen, B. Liu, and X. Qi, "Deep learning for ultrasound image caption generation based on object detection," *Neurocomputing*, vol. 392, pp. 132–141, 2020.

[25] E. Gebhardt and M. Wolf, "Camel dataset for visual and thermal infrared multiple object detection and tracking," in *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pp. 1–6, IEEE, 2018.

[26] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick, "Microsoft coco captions: Data collection and evaluation server. arxiv 2015," *arXiv preprint arXiv:1504.00325*, 2015.

[27] D. Pei, M. Jing, H. Liu, F. Sun, and L. Jiang, "A fast retinanet fusion framework for multi-spectral pedestrian detection," *Infrared Physics & Technology*, vol. 105, p. 103178, 2020.

[28] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

[29] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, pp. 74–81, 2004.

[30] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4566–4575, 2015.