2023 IEEE International Conference on Robotics and Automation (ICRA 2023)
May 29 - June 2, 2023. London, UK

Infrared Image Captioning with Wearable Device

Chenjun Gao, Yanzhi Dong, Xiaohu Yuan, Yifei Han and Huaping Liu*

***Abstract*— Wearable devices have garnered widespread attention as a mobile solution, and various intelligent modules based on wearable devices are increasingly being integrated. Additionally, image captioning is an important task in computer vision that maps images to text. Existing image captioning achievements are based on high-quality visible images. However, higher target complexity and insufficient light can lead to reduced captioning performance and mistakes. In this paper, we present an infrared image captioning framework designed to solve the problem of invalid visible image captioning in special conditions. Remarkably, we integrate the infrared image captioning model into the wearable device. Volunteers perform offline and real-time environmental analysis tasks in the real world to evaluate the framework's effectiveness in multiple scenarios. The results indicate that both the accuracy of infrared image captioning and the feedback from wearable device users are promising.**

I. INTRODUCTION

The image captioning task aims to transform the visual features of images into high-level semantic information so that the computer can generate captions that are similar to those created by humans. In essence, a good caption should be generated by the computer using image features that produce sentences similar to those generated by humans. The caption should contain core information, such as the main objectives, their surroundings, and the relationships between them. Existing developments in image captioning tasks have brought computer-generated captions closer to human-generated captions in many ways. For example, generated sentences or phrases have good structure and fluency [1][2]. The caption is richer and more discriminable [3], or closer to language standards and personalized expressions based on human language structure [4]. Others use reinforcement learning to generate human-approved captions with higher scores [5]. The key point is that current image captioning tasks focus on how to extract image features more accurately, including details of color and action. However, in poorly lit or light-polluted environments, the performance of all visible image captions is drastically reduced. This is because they have high requirements for the clarity and complexity of the image. In the real world, the accuracy of captions is highly dependent on environmental conditions.

Therefore, our starting point is to fully consider many possible scenarios from the real world to improve the generalization of the image captioning application. We take advantage of infrared devices and infrared images to transfer the visible image captioning task to infrared scenes. When the light intensity is below a certain threshold or there is light pollution, the main target may be lost. At this time, infrared sensors are enabled, which are more sensitive to thermal targets than visible light sensors, and are even better than the human eye. In recent years, various technologies have been developed to transform images into speech and sound, which further assist humans in obtaining environmental information. Inspired by research on mobile wearable devices [6], we deploy visible sensors, infrared sensors, and our infrared image captioning model on wearable devices to implement real-time infrared image captioning. The overview is illustrated in Fig. 1.

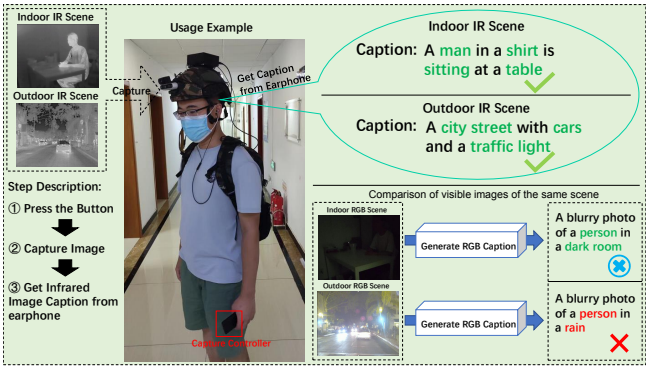

Fig. 1. An demonstration for infrared image captioning on the wearable device. Volunteers carry it for environmental awareness. Press the controller button to capture the current image and output the corresponding caption to the earphone.The comparative examples of necessity are divided into 2 groups: indoor dark environments and outdoor light-polluted environments. The upper left is the captions generated by our infrared image captioning model integrated in the wearable device. Bottom right is visible image with captions generated by traditional image caption pre-trained model. Colored phrases in the caption are keywords. Red means wrong, green means correct, and blue means incomplete information.

In addition, to demonstrate that our model has a certain degree of generalization, we obtained a series of standard scores such as BLEU [7], ROUGE [8], and CIDEr [9] on the validation set of IR-MSCOCO (see Section IV.B). Furthermore, we conducted sample validation on infrared images in existing public datasets FLIR_ADAS_1_3 (see Section IV.C). In conclusion, our infrared image captioning model can cover indoor and outdoor thermal scenes. According to our survey, we are the first to migrate the image captioning task to the infrared scene, which has significantly increased the adaptable range of the image captioning task. Additionally, it is worth mentioning that we have deployed our model on a wearable device, and volunteers wear it to perform completely offline, real-time environmental aware-

All of the authors are with the Department of Computer Science and Technology, Tsinghua University, Beijing, China. Y. Han is also with Department of Mathematical Sciences, Tsinghua University, Beijing, China. Corresponding Author: Huaping Liu (hpliu@tsinghua.edu.cn). This work was supported in part by the National Natural Science Fund for Distinguished Young Scholars under Grant 62025304.

ness tasks in special environments. This also sets the stage for providing assistance to visually impaired individuals. The main contributions are summarized as follows:

- We propose a special scene adaptation scheme for image captioning tasks, which aims to address the problem of information acquisition in dark and light-polluted environments.
- We propose an effective framework for infrared image captioning, and based on this, we develop an infrared image captioning model using deep learning. The model can capture information and relationships between targets from the infrared image and map the infrared image to text.
- We modularize and deploy this model on a mobile wearable device. It can convert environmental visual information into speech, improve people's mobility and perception in special environments. Further, it can also assist visually impaired people in environmental awareness. We have conducted extensive experiments in the real world to demonstrate the validity of our schemes and models.

II. Related Work

A. Image Captioning

Most modern approaches to image captioning utilize a combination of convolutional neural networks (CNN) and recurrent neural networks (RNN), as introduced in [10] and [11]. Since then, many studies have explored and refined this encoder-decoder structure, including attention-based models, such as [2], [12], and [13]. There are also methods to improve captioning performance, such as incorporating reinforcement learning during training, as seen in [14], [15], and [16]. In addition, [3] works to make captions more discriminable, while [4] and [17] focus on improving the richness and diversity of captions by incorporating scene graphs that include the correspondence between object nodes and word nodes. Furthermore, [18] highlights the importance of image features in military decision-making and tactical planning, using key information in captions to assist humans in their next actions. Another recent study by [19] focuses on image captioning in bad weather conditions and proposes a new encoder structure to address the issue of image quality degradation caused by poor visibility.

B. Wearable Devices

The superiority of Artificial Intelligence (AI) in human life is reflected in human-computer interaction and assistive functions. For example, [6] integrates a caption system on glasses, and as a wearable device, it is undoubtedly lightweight. This system enhances the user's perception of speech and sound. Similarly, [20] develops a system that aims to help visually impaired people obtain information from their surrounding environment. The system uses a camera embedded in glasses to capture images and detect text on notice boards or signboards. It then translates the text into speech, making it easier for visually impaired individuals to access important information.

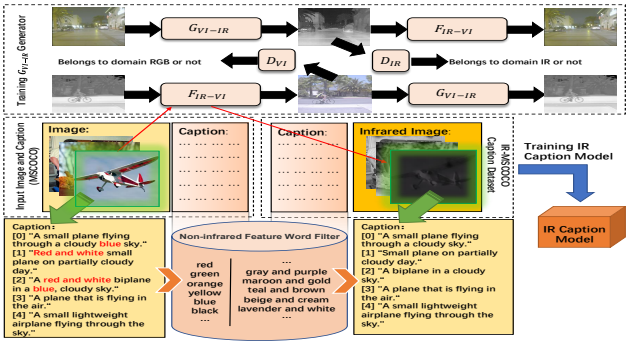

Fig. 2. The architecture of the proposed infrared image captioning framework. It specifically shows the working details of image domain conversion and non-infrared feature filter. The trained infrared image generator converts the visible image in MSCOCO into an infrared image, and the filter filters the non-infrared feature part of the MSCOCO image corresponding to the caption.

C. Application of Infrared Image

In the field of image processing, combining infrared images with deep learning has shown amazing performance. Thermal cameras, with their ability to capture the unique thermal signature of humans, have enabled progress in detection. Initially, human detection and tracking is implemented on infrared image datasets [21], and later, cross-channel pedestrian re-identification is realized, which solves the problem of infrared target recognition under poor lighting conditions or extreme color [22]. A new strategy for robot navigation is proposed by [23], which demonstrates good effectiveness and real-time performance in practical environments. Furthermore, Ref.[24] extends infrared target detection to autonomous driving, with a focus on providing access to information for severe weather and nighttime driving conditions.

III. Structure and Model

The architecture of the infrared image captioning framework is demonstrated in Fig.2. The entire framework consists of three essential modules: image domain translation, infrared image caption dataset and train the infrared image captioning generator. Then we deploy the trained model on a wearable device, and it is carried by volunteers for complete offline and real-time environmental awareness tasks(see section V.B). In the task of electric vehicle (EV) charging inlet detection for autonomous EV charging robots [25], Image-to-Image translation-based data augmentation allows them to achieve higher detection accuracy. So we use "CycleGAN" proposed in [26] to train a generator G_{VI-IR} to translate visible images to infrared. In the section of IR-MSCOCO, we create a non-infrared feature stop word list to filter the text in the caption part of the original dataset MSCOCO, then they are one-to-one corresponded with the above translated infrared images to form a new infrared image caption dataset. At last, for the training part we adopted the method and parameters recommended in [3].

A. Generator G_{VI-IR}

Our ultimate goal of real-world environment awareness is multi-modal image fusion, so visible light images and

infrared images are indispensable. In addition, considering the time cost of data acquisition, we use "CycleGAN" to learn real sensors from unmatched images, so as to achieve bidirectional transfer. In order to adapt our infrared image captioning framework, this section focuses on the generator G_{VI-IR}.

We consider the collection of visible images as the visible light domain X (Domain-VI) and the collection of infrared images as the infrared domain Y (Domain-IR), x and y are the original images in their domains, G_{VI-IR} is the mapping function from Domain-VI to Domain-IR. Similarly, the function from Domain-IR to Domain-VI is represented by F_{IR-VI}. At the same time, G_{VI-IR} and F_{IR-VI} correspond to two discriminators D_{VI} and D_{IR}, which are used to distinguish between real images and generated images as much as possible. So the mapping G_{VI-IR} adversarial loss is defined as follows:

$$\begin{aligned}&\mathscr{L}_{\mathrm{GAN}}\left(G_{VI-IR}, D_{IR}, X, Y\right)=\mathbb{E}_{y \sim p_{\text{IR-image }}(y)}\left[\log D_Y(y)\right] \\ &+\mathbb{E}_{x \sim p_{\text{VI-image }}(x)}\left[\log \left(1-D_Y(G(x))\right)\right]\end{aligned} \tag{1}$$

In order to ensure the unity of the mapping, there is a loss of cycle consistency:

$$\begin{aligned}&\mathscr{L}_{\mathrm{cyc}}\left(G_{VI-IR}, F_{IR-VI}\right)=\mathbb{E}_{x \sim p_{\text{IR-image }}(x)}\left[\|F(G(x))-x\|_1\right] \\ &+\mathbb{E}_{y \sim p_{\text{VI-image }}(y)}\left[\|G(F(y))-y\|_1\right]\end{aligned} \tag{2}$$

The overall loss function is the sum of the adversarial loss and the cycle consistency loss of 2 generators:

$$\begin{aligned}&\mathscr{L}\left(G_{VI-IR}, F_{IR-VI}, D_{IR}, D_{VI}\right)=\mathscr{L}_{\mathrm{GAN}}\left(G_{VI-IR}, D_{IR}, X, Y\right) \\ &+\mathscr{L}_{\mathrm{GAN}}\left(F_{IR-VI}, D_{VI}, Y, X\right)+\lambda \mathscr{L}_{\mathrm{cyc}}\left(G_{VI-IR}, F_{IR-VI}\right)\end{aligned} \tag{3}$$

where λ is the weight ratio of cycle consistency loss and adversarial loss.

We use the FLIR_ADAS_1_3 dataset as training set, which contains a total of 14,452 visible images and 14,452 infrared images, of which 10,228 images are from short videos, and 4,224 images are from videos lasting 144 seconds, and all images are continuous. In order to ensure the discontinuity and diversity of the images, we take 1 frame every 10 frames in the visible and infrared domain, and obtain a total of 1445 images in the visible domain and 1445 in the infrared domain, hereinafter referred to as the visible-infrared image pair. Then we use these visible-infrared image pair to train the generator G_{VI-IR}.

B. Structure of IR-MSCOCO Image Caption Dataset

In the field of Natural Language Processing (NLP), building NLP models and analyzing textual data is essential processes. Where "stop words" help us filter out things that don't have much value in the meaning of the text, it's not a hard rule, but depends on the task we're doing. So we are inspired by it. We add basic NLP functions to the framework, and creat a non-infrared feature stop word list, which can well connect the non-infrared feature filter. In the image caption dataset MSCOCO, each image corresponds to 5 artificially generated captions. There is no doubt that these captions correspond to visible images, not infrared images. Comparing the difference between them, infrared images are sensitive to thermal targets. In the field of image processing, one of the most classic features of infrared image is that the color channel is single, that is, infrared image do not include color features, so all words in captions about color belong to non-infrared features. We mark these color words as non-infrared features, and create a non-infrared feature filter based on the stop word list to filter out all inappropriate captions in the original dataset. An example of the filter is showed in Fig.2. From the perspective of NLP, its essence is a stop word list that contains all the color words.

We aggregate all 123,287 images in MSCOCO and generate corresponding infrared images through the generator G_{VI-IR} mentioned in section III.A. Furthermore, we also feed the 5-sentence caption corresponding to each image into the non-infrared feature filter. The generated images and filtered captions are integrated to make the dataset IR-MSCOCO for training infrared image captioning models. So far, our image domain transformation framework and the non-infrared feature filter work well together and successfully achieve the desired goal. This dataset is the same as the original one, which contains a total of 123,287 images, and each image corresponds to 5 sentences.

C. Infrared Image Caption Net

To make the caption results more diverse, we introduce an attention-based LSTM to generate captions. The CNN is used to obtain the feature vector of the image, then the RNN-based text encoder is used to obtain the feature vector of the caption, and the attention mechanism weights the region when generating the caption. We chose this image caption net because it achieves relatively high scores and highly adaptable in the infrared image scene task, then we retrain the network that has originally been trained on MSCOCO dataset [27] with the aforementioned IR-MSCOCO dataset. In addition, we attempted to incorporate reinforcement learning into training the infrared image captioning model in [28], but found that its generalization and language conciseness were not good enough for wearable device application requirements. Therefore, the model trained without reinforcement learning was retained in the device.

IV. Experiments and Results

The goal of our experiments is to evaluate the performance and effectiveness of the infrared image captioning model derived from our framework, as well as to assess the actual usage effect of the deployed wearable devices. As mentioned in the overview, our motivation for introducing this goal is to solve the problem of environmental awareness in special scenarios through image captioning. Given our focus on wearable devices, it will be essential to not only ensure accuracy on existing datasets but also to assess the device's usability in real-world settings. This will enable real-time environmental awareness by individuals in the real world.

A. Dataset

We train and evaluate on the IR-MSCOCO dataset mentioned in Section III in which a total of 123,287 images participate in training and 40,504 images are used as validation

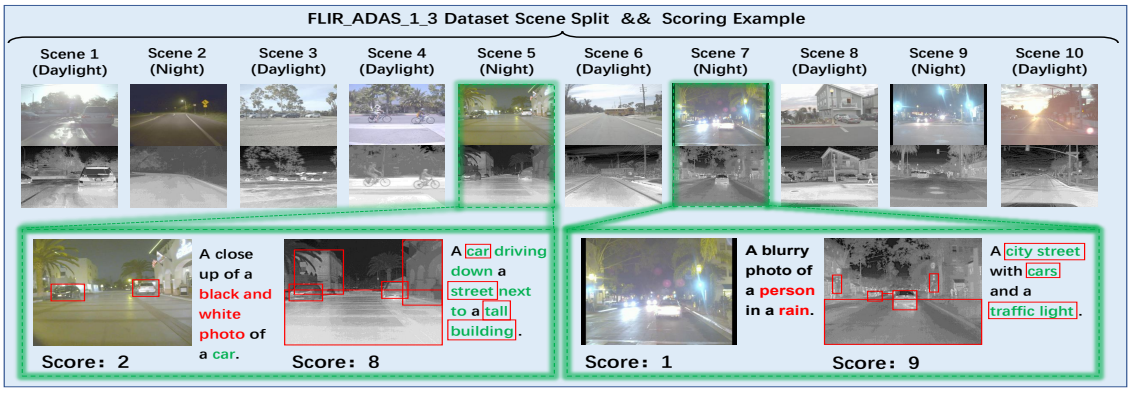

Fig. 3. The landmark images of different scenarios and scoring example.

set to get the score. In order to better reflect real-world scenarios and practicability, we analyze the FLIR_ADAS_1_3 dataset as a cohesive dataset, allowing volunteers to compare the captions generated by our model with the information by human observation of the images to verify the effectiveness.

B. Experimental Design

1) Score: We follow the general quantification metrics of Image Caption, which is a set of standard metrics that have been proposed for evaluating captions. It is down by means of metrics borrowed from machine translation, such as BLEU, ROUGE and CIDEr.

2) Cohesion with real world: We propose this framework is not to achieve higher scores, our goal is to apply it to assist humans in real-world environmental awareness. Therefore, we use the existing dataset FLIR_ADAS_1_3 to link up the real world. The specific evaluation method is shown in section IV.D, and the results are shown in section V.A.

3) Real world tasks for volunteer: Because our infrared image captioning model also has relatively good results (Table I), we further deploy the model on wearable devices for real-world testing by volunteers. We design several scenarios for test, focusing on a dark indoor scene and an outdoor scene in a night with light pollution. At the same time, we also test the Acc(S) and real-time performance of the equipment. The evaluation is based on the validity of the volunteers' real usage feelings and perception of environmental information.

C. Performance on IR-MSCOCO Validation Set

The test set in the MSCOCO dataset does not give a corresponding caption, so we follow the evaluation method of the Image Caption task based on the MSCOCO dataset. We obtain the scores of the infrared image captioning model on the IR-MSCOCO validation set. Because infrared images can also be used as input to generate captions with the visible image captioning model, so we conduct a cross-experiment to illustrate the necessity of our model. The RGB Captioning Model(Visible image captioning model) is taken from the pre-trained model in [3], which has the best score on MSCOCO, and the IR Captioning Model(infrared image captioning model) represents our infrared model. We report the score of the model on validation set in Tab.1. By comparing the second and third rows, in terms of scores, the performance of the RGB Captioning Model on the infrared images is not as good as the IR one. Also, it is very strange to feed infrared images to the RGB image captioning model, and it also gets very bad results.

TABLE I
SCORE OF EACH MODEL'S PERFORMANCE

Methods						
Image Eval Type	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE	CIDRr
RGB Captioning Model for RGB Images	0.778	0.621	0.478	0.363	0.575	1.152
RGB Captioning Model for IR Images	0.522	0.328	0.200	0.126	0.388	0.388
IR Captioning Model for IR Images	**0.653**	**0.469**	**0.328**	**0.231**	**0.481**	**0.673**

D. Performance on FLIR_ADAS_1_3 Dataset

The FLIR_ADAS_1_3 dataset has no human captions, so the Acc(S) cannot be reflected in the form of scores. This also happens to provide us with a new way to test the effectiveness of the infrared image captioning model. According to different video shooting locations and times, we divide the dataset into 10 scenes, and randomly select 10 images from each scene, for a total of 100 images. Each image generates a caption, and then we define a sentence accuracy $\mathrm{Acc}(S), S \in [1, 10]$ and let the volunteers evaluate it, the highest score is 10, and the lowest is 1. We show the landmark images of different scenarios and the scoring process in Fig.3.

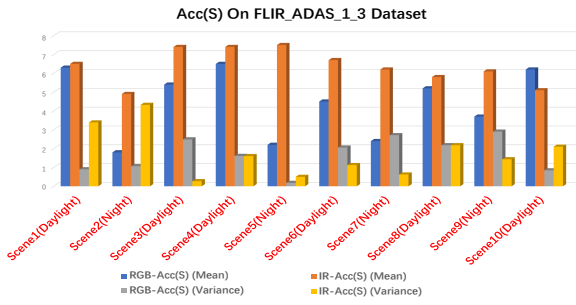

Fig. 4. Acc(S) On FLIR_ADAS_1_3 dataset.

To show the results more evenly, we average the score every 10 images, that is, every scene. In addition, we also perform the same evaluation on the visible light image corresponding to the infrared image. The results are shown

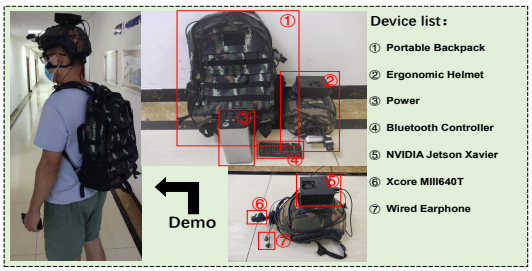

Fig. 5. Wearable display.

in Fig.4. For the daytime evaluations, the results of the two evaluations are very similar. And it can be visualized that the infrared image captions are all higher than the visible image captions in the Night Scenes, and the stability can be seen by the variance.

E. Wearable Device and Task

1) Introduction of Wearable Device: The wearable device is an ergonomic helmet. The included accessories are Sourlor 68000mAh portable power, Sony in-ear USB wired earphone, Xcore MIII640T infrared camera, NVIDIA Jetson Xavier, Ultra MINI Bluetooth keyboard. We embed the ITS (Image-To-Speech) module and deploy our infrared image captioning model to NVIDIA Jetson Xavier. The user controls the infrared camera through the Bluetooth remote button to obtain the infrared image of the current scene and output its voice caption to wired earphone. The device can run completely offline, has excellent stability, and avoids the communication dependence problem of real-time backhaul processing in the cloud. In addition, after our tests, the shortest time from the volunteer pressing the button to hearing the sound is 1.7 seconds, the longest is 2.1 seconds, and the average is 1.8 seconds. In other words, it has excellent real-time and stability.

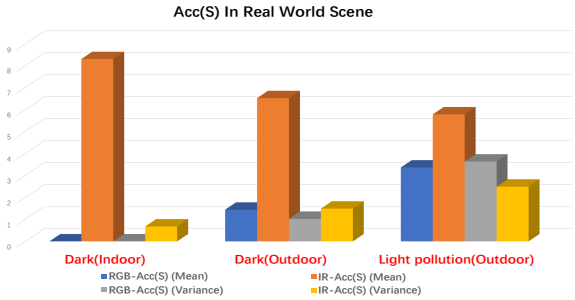

Fig. 6. Acc(S) In Real World Scene.

2) Task Introduction and Scene: We package the keyframes of the volunteers' actual tasks into a collection of scene test images. There are 2 major categories and 2 sub-categories. The major categories are indoor scenes and outdoor scenes, and 2 sub-categories include darkness and light pollution, 107 images in total. The accuracy test is consistent with the method mentioned in section IV.D, and the results are shown in Fig.6. It is clear that the evaluation scores for the infrared image captions are all higher than those for the visible image captions. In addition, the actual task demonstration graphs(Fig.8) that we show in section V are both included in the above test graph collection.

V. Qualitative Result and Analysis

We will present the qualitative results in three parts. Fig.7 is a comparison of the sampled images of the dataset FLIR_ADAS_1_3. Fig.8 shows the results of infrared images captured by volunteers in a real-world task with wearable devices and the caption results (in this part, the caption results of the infrared images have been converted into speech through ITS module and output to the ear). We also found some possible problems in the comprehensive test, in some scenes and shooting angles, wrong captions will be generated (Fig.9).

A. Part of Results on FLIR_ADAS_1_3

The scene shown in the Fig.7 contains light pollution scenarios during the day and night, and we highlight the important target that can be observed by naked eye. The left side is a comparison diagram result of each captioning model of each image. Obviously, in this case, the caption of the visible image is completely inaccurate, while the caption of the infrared image is relatively accurate and has a certain sentence complexity. In addition, in order to ensure the integrity of the experiment, we also put the same infrared image into the visible image captioning model, the result is also inaccurate. The redundancy of words about the color channel in the caption corresponds to the important role of the non-infrared feature filter in the previous section.

B. Part of Comparison and Display of Real World Tasks

As illustrated in Fig.8, we divide the task into outdoor scene and indoor scene. Due to the particularity of infrared application scenarios, we all choose to perform infrared image captioning tasks at night and in the darkroom. We capture both visible images and infrared images while performing the

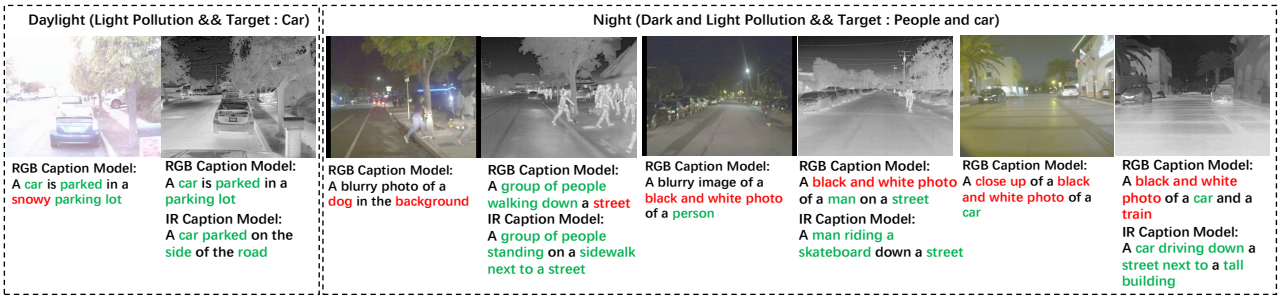

Fig. 7. The scene is RoadSence(Daylight and Night) of FLIR_ADAS_1_3 dataset. We show part of the sampled images.

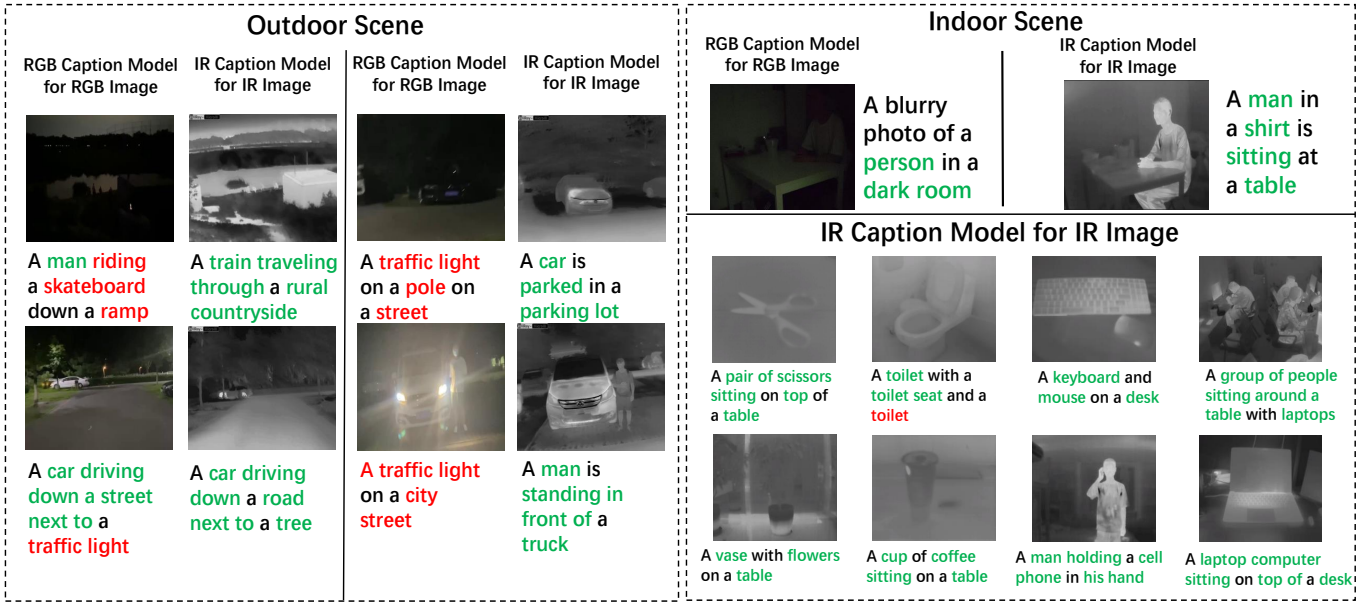

Fig. 8. Outdoor and indoor scenes comparison of image captioning.

task. There are some contrasting observable objects in the visible images of the outdoor environment, while the indoor environments are all dark. For indoors, after showing a set of comparison images, we only show infrared images for the rest. We no longer show the results of the visible image captioning model on infrared images.

C. Bias and Analysis

We encounter many examples of inaccuracy or incoherence in volunteer tasks, as shown in Fig.9. We detail the inaccurate information and preliminary reasons for the error in the "Mistake". Sometimes it cannot distinguish targets with similar characteristics, while humans cannot directly judge targets by infrared images either. Humans need to combine the surrounding environment to make judgments. We believe that the accuracy of caption can be further improved by increasing the image resolution and correlating the target environment information.

VI. CONCLUSIONS

In this paper, we propose an adaptation scheme for image captioning tasks in special scenarios using wearable devices, based on their superiority in assisting humans with environmental awareness tasks. We establish a usable framework that includes three basic modules: image domain translation, infrared image dataset, and infrared captioning net. The framework is evaluated for connectivity on the FLIR_ADAS_1_3 dataset, and then integrated into a wearable device to perform fully offline, real-time tasks in real-world indoor and outdoor scenes. Comprehensive experiments demonstrate the effectiveness of our wearable device and framework. However, our work has certain limitations: it only considers the respective characteristics of visible images and infrared images in a specific environment and does not fully utilize their advantages simultaneously. In future work, we will continue to use the advantages of bidirectional transfer with "CycleGAN" and make full use of the generator F_{IR-VI} to further fuse visible and infrared images, concentrating their advantages to achieve the preservation of important target details of dual images. This will enable the wearable device to generate more accurate, comprehensive, and rich multi-model captions.

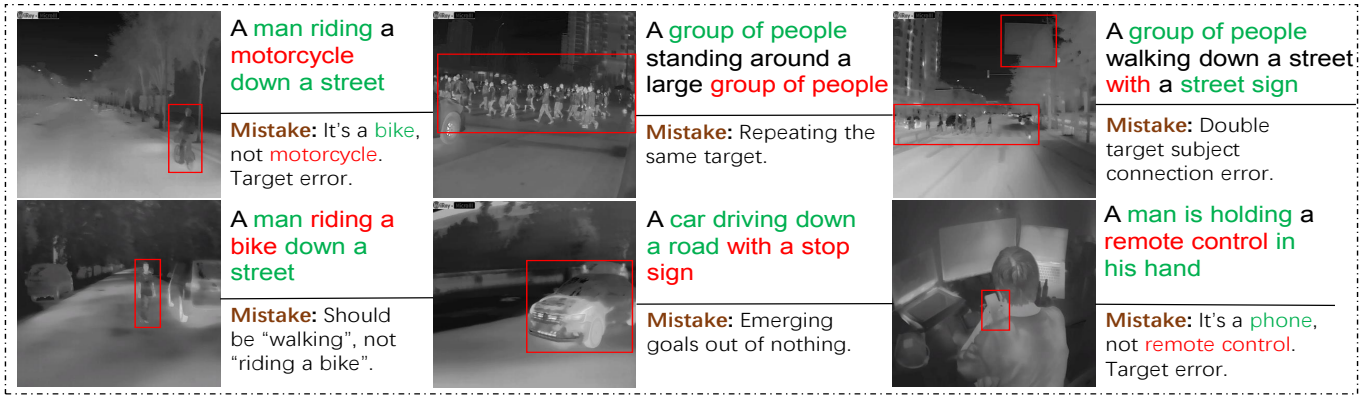

Fig. 9. Examples of some inaccurate image captions.

REFERENCES

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

[2] 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*, 2018, pp. 6077–6086.

[3] 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*, 2018, pp. 6964–6974.

[4] 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*, 2020, pp. 9962–9971.

[5] 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*, 2017, pp. 7008–7024.

[6] 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*, 2020, pp. 1108–1120.

[7] 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*, 2002, pp. 311–318.

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

[9] 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*, 2015, pp. 4566–4575.

[10] 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*, 2015, pp. 3156–3164.

[11] 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*, 2015, pp. 3128–3137.

[12] 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*. PMLR, 2015, pp. 2048–2057.

[13] 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*, 2017, pp. 375–383.

[14] M. Ranzato, S. Chopra, M. Auli, and W. Zaremba, "Sequence level training with recurrent neural networks," *arXiv preprint arXiv:1511.06732*, 2015.

[15] S. Liu, Z. Zhu, N. Ye, S. Guadarrama, and K. Murphy, "Improved image captioning via policy gradient optimization of spider," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 873–881.

[16] L. Zhang, F. Sung, F. Liu, T. Xiang, S. Gong, Y. Yang, and T. M. Hospedales, "Actor-critic sequence training for image captioning," *arXiv preprint arXiv:1706.09601*, 2017.

[17] X. Yang, H. Zhang, and J. Cai, "Auto-encoding and distilling scene graphs for image captioning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[18] 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)*. IEEE, 2021, pp. 1–8.

[19] C.-H. Son and P.-H. Ye, "New encoder learning for captioning heavy rain images via semantic visual feature matching," *Journal of Imaging Science and Technology*, vol. 65, no. 5, pp. 50 402–1, 2021.

[20] C. Rane, A. Lashkare, A. Karande, and Y. Rao, "Image captioning based smart navigation system for visually impaired," in *2021 International Conference on Communication information and Computing Technology (ICCICT)*. IEEE, 2021, pp. 1–5.

[21] J. Portmann, S. Lynen, M. Chli, and R. Siegwart, "People detection and tracking from aerial thermal views," in *2014 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2014, pp. 1794–1800.

[22] X. Xiang, N. Lv, M. Zhai, R. Abdeen, and A. El Saddik, "Dual-path part-level method for visible–infrared person re-identification," *Neural Processing Letters*, vol. 52, no. 1, pp. 313–328, 2020.

[23] M. F. Xaud, A. C. Leite, and P. J. From, "Thermal image based navigation system for skid-steering mobile robots in sugarcane crops," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 1808–1814.

[24] F. Munir, S. Azam, and M. Jeon, "Sstn: Self-supervised domain adaptation thermal object detection for autonomous driving," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 206–213.

[25] Y. Bang, Y. Lee, and B. Kang, "Image-to-image translation-based data augmentation for robust ev charging inlet detection," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 3726–3733, 2022.

[26] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.

[27] 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 preprint arXiv:1504.00325*, 2015.

[28] C. Gao, G. Bian, Y. Dong, X. Yuan, and H. Liu, "Infrared image captioning based on unsupervised learning and reinforcement learning," in *2022 International Conference on Automation, Robotics and Computer Engineering (ICARCE)*. IEEE, 2022, pp. 1–4.