2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

A Lightweight De-confounding Transformer for Image Captioning in Wearable Assistive Navigation Device

Zhengcai Cao*, *Senior Member, IEEE*, Ji Xia, Yinbin Shi, *Member, IEEE*
and MengChu Zhou, *Fellow, IEEE*

***Abstract*— Image captioning is a multi-modal task that enables the transformation from scene images to natural language, providing valuable insights for visually impaired individuals to understand their environment. Therefore, its application to wearable navigation devices for visually impaired individuals holds immense potential. However, in practical applications, confusion between scene visuals and semantics, coupled with model complexity, often leads to performance degradation, resulting in inaccurate environmental interpretation. In light of this, we introduce a Lightweight De-confounding Transformer Network (LDTNet) for image captioning equipped with a Causal Adjustment module to eliminate confounders. Moreover, we design a Suppression Gate Unit that efficiently integrates fine-grained information from shallow features, while reducing the number of network layers to have a lightweight model. Experimental results demonstrate that our approach not only addresses the visual-semantic confusion issue effectively but also improves the response speed of wearable devices in comparison with the state of the art. Twenty volunteers are recruited to evaluate LDTNet's efficacy in real-world settings in terms of both response speed and generated outputs by wearing the resulting assistive navigation devices. The outcomes well show its outstanding performance and great potential for visualy impaired individuals to use.**

I. INTRODUCTION

Image captioning is an interdisciplinary task, encompassing the realms of computer vision (CV) and natural language processing (NLP). It is required to establish connections between the primary objects and the ambient environment in an image, yielding descriptions in human-understandable natural language [1], [2]. Consequently, image captioning has a significant value for giving visually impaired individuals the insight into their surroundings.

A good image caption is expected to be accurate, fluent, and close to human language. Existing image captioning methods emphasize fluent descriptions while highlighting object interactions and detailed content in images [3]. Many image captioning efforts, are dedicated to extracting richer visual features, while overlooking the confounding factors between vision and language. When two objects frequently appear together in scene, a 'shortcut' might form between them, leading to a spurious correlation and consequently incorrect results. For example, in Fig. 1, people usually use umbrellas on rainy days. The frequent co-occurrence of 'umbrella' and 'rainy day' can lead to confusion. When people use umbrellas on sunny days to block sunshine, due to this confusion, the model tends to produce descriptions suggesting rain. In practical applications, due to the concentration of required scenarios, the probability of such confusion increases, causing great discrepancies in the understanding of the environment by visually impaired individuals.

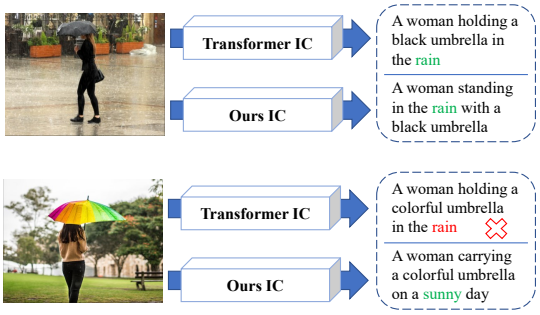


Fig. 1. An example of confounding in image captioning. The visual feature of an 'umbrella' often leads to the generated word 'rain'. In this scenario, 'Transformer Image Captioning' represents standard Transformer image captioning model, whereas 'Ours Image Captioning' refers to our introduced model. Words that are correctly and incorrectly generated are colored in green and red, respectively.

The primary objective of this paper is to address the visual-semantic ambiguities that frequently arise in various scenarios, enhance the model's generalization capabilities, and assist the visually impaired in accurately perceiving their surroundings. We propose a computationally efficient image captioning model tailored to counteract this confusion. Inspired by [4], a two-layer Transformer is adopted in the encoder. We introduce a Suppression Gate Unit (SGU) and leverage skip connections to efficiently integrate fine-grained information from shallow features. Furthermore, we propose a Causal Adjustment (CA) module that introduces an intermediate variable to eliminate confounding factors through front-door adjustment. Starting with the aim to assist visually impaired individuals in accurately understanding their surroundings, we deploy visual sensors and image captioning models in wearable assistive navigation devices for testing and evaluation. The image captioning function adopts a pause-triggered mechanism, providing voice announcements when the visually impaired individual encounters obstacles.

This work is supported in part by the National Natural Science Foundation of China under Grant (92148202, 52175002), and the Beijing Natural Science Foundation (L223019). (Corresponding author: Zhengcai Cao.)
Zhengcai Cao is with State Key Laboratory of Robotics and Systems, Harbin Institute of Technology, Harbin, 150080, China. (caozc@hit.edu.cn)
Ji Xia and Yinbin Shi are with the College of Information Science and Technology, Beijing University of Chemical Technology, Beijing, 100029, China. (e-mail: xiaji001115@163.com and syb2513@163.com).
MengChu Zhou is with the School of Information and Electronic Engineering, Zhejiang Gongshang University, Hangzhou 310018, China (mengchu@gmail.com).

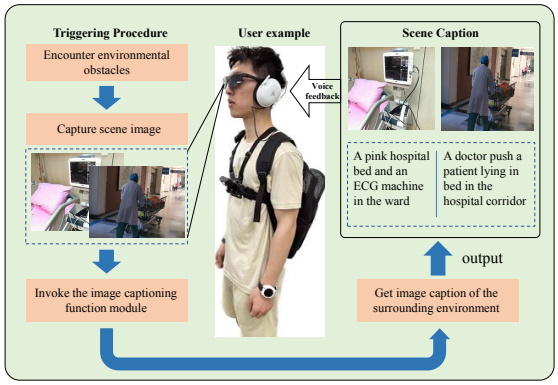


Fig. 2. A demonstration of the image captioning function on the wearable assistive navigation devices. Volunteers wear it to gain insights into their environmental surroundings. When a volunteer's movement pauses, the visual sensor captures the immediate scene, and an auditory description is relayed through the earpiece.

The overview is illustrated in Fig. 2.

To validate its effectiveness, our approach is evaluated on the MSCOCO dataset to obtain a series of image captioning metrics and compared with state-of-the-art methods. Furthermore, we deploy it to wearable assistive navigation devices for the visually impaired. Multiple participants take part in testing. Based on both the scores given by the participants [5] and standard evaluation metrics, we demonstrate that our model can express the contents of a scene rapidly and accurately. Our new contributions to the field of image captioning are as follows:

- We propose a Causal Adjustment (CA) module to address the issue of visual features in image captioning that are prone to confusion. This module can effectively eliminates most confounding factors.
- We propose a Lightweight De-confounding Transformer Network (LDTNet) architecture with an SGU and CA module to achieve excellent performance while ensuring fast response for applications.
- We implement this architecture on wearable assistive navigation devices. A substantial number of real-world trials conducted by various volunteers have well showcased its efficacy of our approach.

II. Related Work

A. Image Captioning

Numerous approaches employ convolutional neural networks (CNN) and recurrent neural networks (RNN) to set up an encoder-decoder architecture for caption generation [6], [7]. Building on this foundation, there has been an increasing focus on highlighting key regions in images by introducing attention mechanisms [8], [9]. With the introduction of visual Transformer [10], [11], recent methods that integrate CNN with Transformer have exhibited their exceptional performance [2], [12]. Further, [13] incorporates task-adaptive vectors in a decoder to mitigate the perturbations influence of visual features when generating non-visual words. Luo *et al.* [14] fuse various visual features and model the interrelations among them to obtain stronger visual representations. Some methods [15], [16] achieve significant performance improvements by adopting reinforcement learning during a model training process.

B. Causal inference

Some researchers have integrated causal inference into neural network models for computer vision tasks, resulting in significant performance improvements. This inclusion of causal inference can be observed in tasks like image classification [17], and semantic segmentation [18]. In the area of image captioning, Yang *et al.* [19] have constructed a Deconfounded Image Captioning (DIC) framework by analyzing various confounding biases. However, they offer a thorough analysis of intertwined visual and semantic factors. Liu *et al.* [20] have designed a structure that counters confounding through backdoor adjustment. Yet, their method demands a meticulous quantitative analysis of confounding factors, making it complex and infeasible for real-world applications. As a novel solution, we introduce a streamlined front-door adjustment approach by incorporating intermediary variables into our proposed architecture to maintain efficiency while eliminating most confounding elements.

III. Proposed Methods

The overall architecture of our proposed LDTNet is shown in Fig. 3. To achieve model lightweighting, we draw inspiration from [4] by employing a 2-layer Transformer structure in the encoder. At the same time, we integrate shallow visual features by using skip connections and SGU to bolster visual representation. In addition, we introduce a CA module into the decoder for generating image captions, thus well addressing the visual-semantic ambiguity issue.

A. Feature Extraction and Encoding

The prevailing captioning methods typically rely on object detection networks to extract region features or utilize deep convolutional layers to obtain grid features. To facilitate model deployment and save computational resources greatly, we fully exploit the advantages of fast grid feature extraction. Although deep convolutional features possess more abstract semantic information, they lack some fine-grained details. Additionally, during a Transformer encoding process, certain fragile visual information may be filtered out. To address this issue, we introduce an additional single convolutional layer to extract shallow features and integrate it into the main encoding pathway through skip connections. Moreover, we construct a SGU in a skip connection process to reduce the weight of shallow features, thereby avoiding an excessive dominance of edge texture information and noise in the visual representation.

SGU learns appropriate suppression weights through a fully connected layer followed by a sigmoid activation function. Taking into account the interdependencies and influences between shallow features and deep ones, we jointly consider

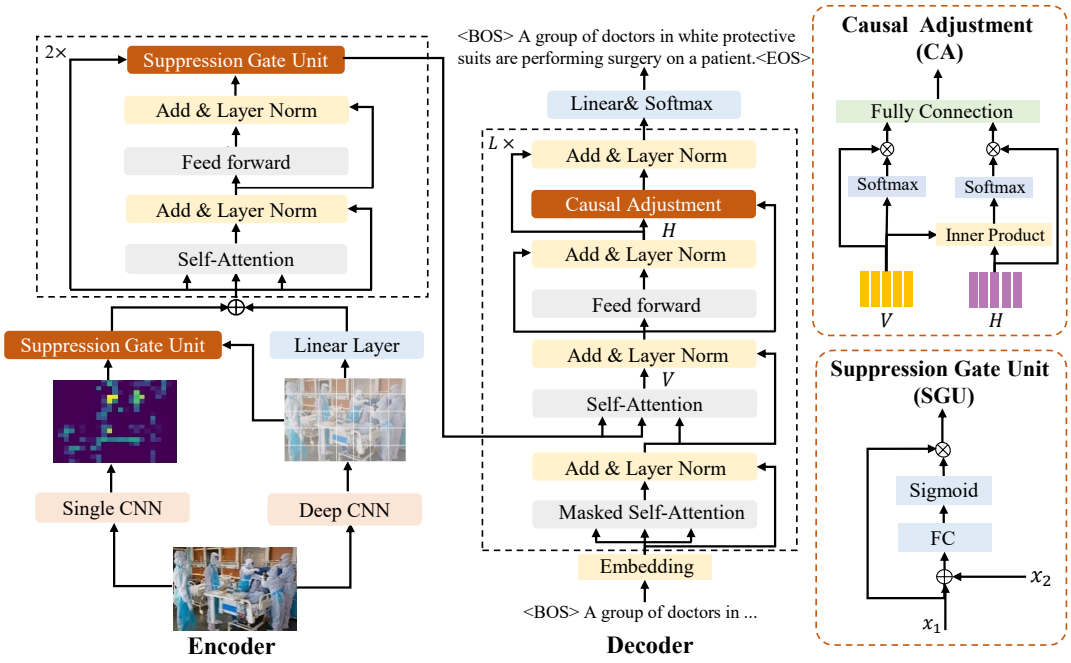


Fig. 3. The overview of our proposed LDTNet. Based on the Transformer encoder-decoder architecture, our approach adds skip connections and SGU to integrate shallow features and backbone ones. A CA module is introduced in the decoder to eliminate visual semantic confusion.

the current layer features X_c and subsequent layer features X_s as inputs. The computation process is:

$$w_s = Sigmoid(FC(X_c + X_s)), \tag{1}$$

$$Z = w_s \odot X_c, \tag{2}$$

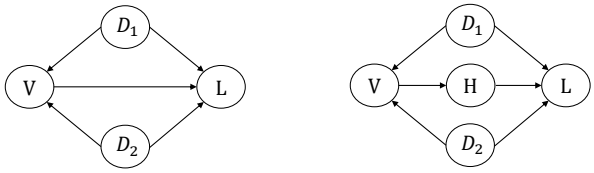
where FC denotes the computation process of a fully connected layer, $Sigmoid$ represents a sigmoid activation function, w_s denotes the suppression weights, and z represents SGU's output. We apply skip connections and SGU between the two Transformers in the encoder to properly leverage the useful information from the preceding layer features.

Compared to our previous work [21], this method abandons the calculation of complex geometric relationships and combines two types of image features more efficiently through the SGU module.

B. De-confounding decoder

To address the visual-semantic confusion issue in image captioning, we propose a CA module. The decoding process sequentially generates textual descriptions over time steps. Throughout this procedure, the visual and language features exhibit a causal relationship, as illustrated in Fig. 4(a). Herein, V denotes visual features, L represents the generated words, and D_1 and D_2 correspond to the visual and linguistic confounders, respectively, with the assumption that D_1 and D_2 are independent.

The front-door adjustment can leverage the information of the mediator variable to indirectly control the effect of confounding variables. This method has been widely recognized in causal inference but is relatively complex



(a) causal confounded (b) front-door adjustment

Fig. 4. (a) shows the Causal diagram of image captioning process. (b) shows the front-door adjustment method to perform causal adjustment on the causal diagram for image captioning.

to implement [19]. Therefore, we put forth a streamlined front-door adjustment technique to mitigate the confounding effect. As depicted in Fig. 4(b), we introduce an intermediate variable H to the causal graph, interpreted as a fusion of visual features with the prior semantic attribute. Subsequently, we compute $P(L|do(V))$ through the causal trajectory $V \rightarrow H \rightarrow L$:

$$P\left(L|do\left(V\right)\right) = \sum_{V} P\left(V\right) \sum_{H} P\left(H|V\right) P\left(L|H,V\right). \tag{3}$$

where $P(L|do(V))$ expresses the probability of producing a sentence L under the causal inference given the observed visual feature V. For computational convenience, we turn to the Normalized Weighted Geometric Mean approximation [22], [23], and employ a Softmax function to determine

the expected value:

$$P\left(L|do\left(V\right)\right) \approx Softmax\left\{FC(\mathbb{E}_V[V], \mathbb{E}_{H|V}[H])\right\} \quad (4)$$

$$\mathbb{E}_V[V] = Softmax(V)V, \quad (5)$$

$$\mathbb{E}_{H|V}[H] = Softmax(VH^T)V, \quad (6)$$

where $\mathbb{E}$ stands for the mathematical expectation and the Softmax-derived estimate represents the probability of each element in the corresponding input. We embed CA module into each Transformer decoding block to generate deconfounded image captions, as shown in Fig. 3.

C. Training Details

Consistent with the training strategies adopted in [2], our image captioning training process is generally organized into two stages [15], [24]. During the pre-training phase, given an image I and a sentence $Y = \{y_1, y_2, ..., y_n\}$, our model can be optimized using the cross-entropy loss:

$$L_{CE}(\theta) = -\sum_{i=1}^{n}(\log(p_\theta(y_i|y_{1:i-1}))), \quad (7)$$

where θ represents the model's parameters. Following the principles of language modeling (LM), the model is trained using the subsequent loss function:

$$l = \frac{1}{n+1}\sum_{i=1}^{n+1} L_{CE}(y_i, p(w_i|I, y_{1:i-1})). \quad (8)$$

For the fine-tuning phase, we employ self-critical sequence training [15]. The objective is to minimize the negative CIDEr-D [25] scores:

$$L_{RL}(\theta) = -E_{w_1:n\sim p_\theta}[r(y_{1:n})], \quad (9)$$

where r denotes the CIDEr-D score function. The gradient expression for a single sample can be approximated as:

$$\nabla_\theta L_{RL}(\theta) \approx -\frac{1}{k}\sum_{i=1}^{k}((r(y^i_{1:T}) - b)\nabla_\theta \log p_\theta(y^i_{1:T})). \quad (10)$$

In this equation, $y^i_{1:n}$ is the i-th sequence, k represents the number of sequences, and b corresponds to the average reward obtained from the sampled sequences.

IV. Experiments and Results

We conduct extensive experiments to validate the effectiveness of the proposed LDTNet for our wearable assistive navigation device. LDTNet is deployed on the actual wearable device for testing and evaluation. In the experiments, our primary focus is on assessing whether it can effectively assist visually impaired individuals (volunteers) in gaining a better understanding of their surroundings. We specifically evaluate the ability to promptly respond and provide feedback when the captioning function is triggered.

A. Datasets

To compare with the state-of-the-art methods, we use the Microsoft COCO dataset, which consists of 123,287 images, each with 5 captions. We employ the Karpathy split [26], where 5,000 images are set aside for validation, another 5,000 for testing, and the remaining images for training. Moreover, to further demonstrate the practical utility of LDTNet, we have manually annotated a dataset (IGS_H_1) tailored for indoor guidance scenarios for the visually impaired. This dataset primarily contains 3,000 images (2,400 for training, 300 for validation, and 300 for testing). Volunteers equipped with devices in real-world situations assess the captions generated by LDTNet trained on this custom dataset based on their observed information, validate the efficacy of our approach.

B. Experiment Design

Evaluation metrics: We follow the standard metrics for image captioning to evaluate the semantic correctness and fluency of captions, including BLEU [27], ROUGE [28], METEOR [29], CIDEr [25], and SPICE [30]. In the wearable device testing tasks, we measure the response time from the moment triggering the image captioning function to the time when volunteers receive voice feedback for validating LDTNet's real-time capabilities. We score the model based on the users' actual experience and objective evaluations from other individuals to assess the accuracy of the model in generating captions.

Implementation Details: We employ a standard Transformer encoder-decoder architecture as a foundation, adhering to some configurations specified in [4]. Specifically, we use two encoder layers and four decoder ones. For each Transformer, the input and output dimensions are set to 512, with the number of heads set to 8. A dropout probability of 0.1 is applied after each attention layer and position-wise feed-forward layer. For image feature extraction, we utilize an optimized ResNeXt152 [31] to arrange the features into a 7×7 grid. The grid features are then flattened and reshaped into a 49×512 two-dimensional tensor, serving as the input for the encoder. All models used in our experiments are trained or evaluated on a single NVIDIA RTX3090 24GB GPU. We employ the Adam optimizer for training our model with $\beta_1 = 0.9$ and $\beta_2 = 0.98$. The learning rate follows an epoch decay schedule [2].

C. Ablation Analysis

We first conduct ablation experiments to validate the effectiveness of the newly proposed SGU and CA in improving LDTNet's performance. The experiments start with a baseline model that consists of ResNeXt152 and Transformer without any additional modules. Then, we add shallow feature skip connections (SC), SGU module, and CA module separately for experimentation. Finally, these modules together form our LDTNet. The results are shown in Table I. Noticeably, the direct inclusion of shallow features does not yield a significant improvement. In the process of continuous skip connections, the textural details and inherent noise

TABLE I

ABLATION STUDY OF OUR PROPOSED MODULE ON THE MSCOCO KARPATHY TEST SPLIT

SC	SGU	CA	B@1	B@4	M	R	C	S
✗	✗	✗	81.1	39.1	29.3	58.9	132.8	23.1
✓	✗	✗	81.4	39.3	29.3	58.9	132.9	23.2
✗	✓	✗	81.4	39.5	29.5	59.3	133.4	23.2
✗	✗	✓	81.5	39.4	29.4	59.2	133.6	23.3
✗	✓	✓	81.9	39.9	29.7	59.4	134.7	23.3
✓	✗	✓	81.8	39.9	29.6	59.3	134.4	23.2
✓	✓	✗	81.7	39.8	29.5	59.1	134.6	23.3
✓	✓	✓	**82.3**	**40.4**	**29.9**	**59.5**	**135.9**	**23.5**

TABLE II

PERFORMANCE COMPARISION WITH THE STATE OF THE ART ON THE MSCOCO KARPATHY TEST SPLIT

Model	Params	B@1	B@4	M	R	C	S
Up-Down [8]	52.1M	79.8	36.3	27.7	56.9	120.1	21.4
SGAE [32]	125.7M	80.8	38.4	28.4	58.6	127.8	22.1
AoANet [24]	87.4M	80.2	38.9	29.2	58.8	129.8	22.4
SMArT [4]	-	80.4	38.1	28.8	58.2	129.7	22.2
M^2Transformer [33]	38.4M	80.8	39.1	29.2	58.6	131.2	22.6
SCD-Net [34]	-	81.3	39.4	29.2	59.1	131.6	23.0
X-Transformer [7]	137.5M	80.9	39.7	29.5	59.1	132.8	23.4
DLCT [14]	-	81.4	39.8	29.5	59.1	133.8	23.0
RSTNet [2]	-	81.8	40.1	29.8	**59.5**	135.6	23.3
LDTNet(ours)	41.6M	**82.3**	**40.4**	**29.9**	**59.5**	**135.9**	**23.5**

from these shallow features tend to overshadow the semantic information of the main features. By introducing SGU, the model can adaptively weigh these features, thus effectively mitigating the noise and incorporating more fine-grained and useful features into LDTNet, and eventually leading to an enhancement in LDTNet's performance. After adopting CA, its performance is further improved. This improvement can be attributed to CA's ability to eliminate biases caused by confounding factors between visual information and semantic one during the decoding phase. After we combine these modules together, we achieve more significant improvement as shown in Table I.

D. Comparison with the state of the art

We conduct performance comparisons between LDTNet and its competitive peers, *i.e.* Up-Down [8], SGAE [32], AoANet [24], SMArT [4], M^2Transformer [33], SCD-Net [34], X-Transformer [7], DLCT [14], and RSTNet [2]. Up-Down is the most classic CNN-LSTM network that utilizes region features, while SMArT is a lightweight Transformer network. SGAE proposes a graph convolutional network that achieves excellent results. AoANet, M^2Transformer, and X-Transformer are recently widely compared high-performance models. SCD-Net, a recently proposed model, introduces a novel diffusion modeling paradigm. DLCT combine the advantages of multi-visual features to achieve significant performance, but consuming more computational resources. RSTNet achieves state-of-the-art performance by utilizing Grid-Augmented module and Adaptive-Attention module. As observed in Table II, our model achieves significant performance advantages with a relatively small number of parameters.

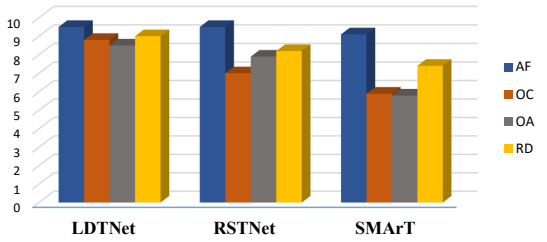


Fig. 5. Comparison of average scores for each evaluation criterion.

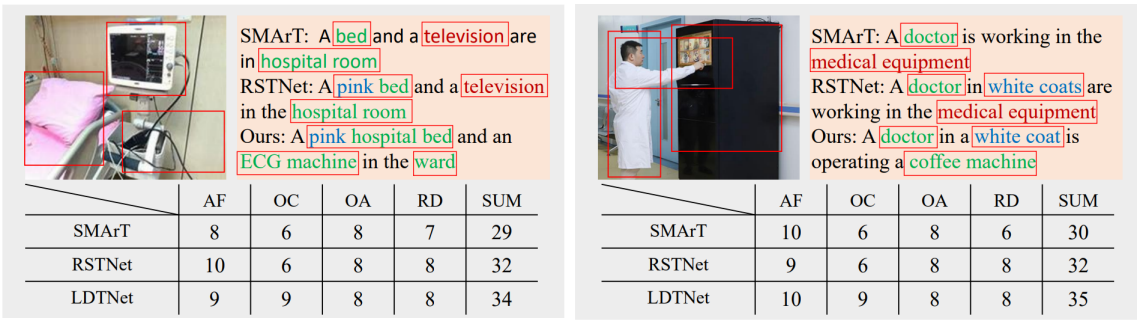
E. Deployment and Evaluation on Wearable Device

To further validate LDTNet's performance in practical applications, we deploy it into a wearable assistive navigation device and build an image captioning function module specifically designed to serve visually impaired individuals.

1) Wearable assistive navigation device: Our wearable assistive navigation device adopts an industrial computer (NVIDIA Jetson AGX ORIN64GB) serving as the core processor and placing in a heat dissipation backpack along with the power supply. The input accessories include video glasses and a chest-mounted monocular camera, which communicate via a USB serial port. The output accessories consist of two vibration wristbands and a bone conduction headphone, which communicate via Bluetooth. An Inertial Measurement Unit (IMU) on the shoulder is used to determine the volunteer's motion status.

We embed LDTNet into the industrial computer and trigger the image captioning function based on a pause-triggered approach in accordance with the obstacle avoidance system. The system segments an input image from the chest camera into passable areas and obstacles, providing feedback through the vibration wristbands. When the volunteer encounters an obstacle preventing them proceeding or when the volunteer autonomously pauses for more than a given threshold, the image captioning function is triggered. The real-time image captured by the video glasses is used as input, and the system generates a scene description that is converted into speech and fed back to the volunteer through the headphones.

2) Performance on real world scenes: In a real hospital indoor setting, volunteers with wearable assistance navigation devices test their performance. In this experiment, a total of 20 volunteers participate, among them 10 are sighted individuals wearing blindfolds (5 males and 5 females, with an average age of 31 years), 7 are partially sighted individuals with vision less than sixty percent of normal (6 males and 4 females, with an average age of 37 years). We compare our model with the high-performance RSTNet and lightweight SMArT. Notably, to make the them more suitable for the real-world experimental scene, all models are pre-trained on MS-COCO and then fine-tuned on our custom dataset IGS_H_1.



	AF	OC	OA	RD	SUM
SMArT	8	6	8	7	29
RSTNet	10	6	8	8	32
LDTNet	9	9	8	8	34

	AF	OC	OA	RD	SUM
SMArT	10	6	8	6	30
RSTNet	9	6	8	8	32
LDTNet	10	9	8	8	35

Fig. 6. Examples of detailed evaluations for real-world scene images. Red text indicates incorrect object descriptions, green signifies correct object descriptions, and blue denotes partial detail descriptions.

Since real-world scene images do not have standard manual captions, we refer to the methods from [5] and further refine our evaluation criteria. For each image, volunteers rate the navigation devices in terms of the following metrics: sentence grammatical accuracy and fluency (AF); primary object coverage (OC); object association (OA); and richness of details (RD). The score for each criterion falls into [0,10]. We take the average score given by volunteers as the final rating for each image. It is worth noting that when the model produces incorrect object descriptions due to visual-semantic confusion, a very low object coverage score is given. We evaluate 50 real-scene images and compare the average scores of different models. As shown in Fig. 5. we conclude that our model garner higher evaluations.

3) Qualitative analysis: To provide a deeper insight and analysis into the experimental outcomes, two representative examples with detailed results are presented in Fig. 6. The variance in model outputs is evident in terms of object coverage and the richness of details. In Fig. 6, both SMArT and RSTNet misidentify the ECG machine as a television and interpret the coffee machine as medical equipment. A plausible explanation for this misidentification can be attributed to the frequent concurrent presence of televisions with beds and physicians with medical apparatus in data samples, leading to a spurious correlation between them. Our proposed model, including CA module, effectively diminishes this association, thereby preventing such confusion and misclassification.

4) Computational time analysis: We assess the response time of LDTNet in wearable devices and benchmark it against SMArT and RSTNet. SMArT stands as one of the most representative lightweight Transformer models in recent years with strong performance. RSTNet, on the other hand, combines deep convolution with Transformers, resulting in high performance. Our experimental evaluation encompasses multiple scenarios, measuring the time from triggering the image captioning function to when the device begins issuing auditory descriptions. Fig. 7 illustrates the response times across 15 real-world scenarios for the different models.

A notable observation is that LDTNet, with an average response time of 1.10s, outperforms RSTNet, which has an average response time of 1.53s. This is due to the deeper network architecture of RSTNet, which entails higher computational demands. Compared to SMArT, which has an average response time of 1.01s, our model's response time is slightly longer but remains highly competitive. While our architecture incorporates additional modules and two decoder layers, it avoids the computational overhead of a dedicated object detection network by leveraging ResNeXt152 for feature extraction. Moreover, given that the caption quality of SMArT (CIDEr 129.7) is lower than ours (CIDEr 135.9), LDTNet achieves a significant performance boost with minimal computational overhead.

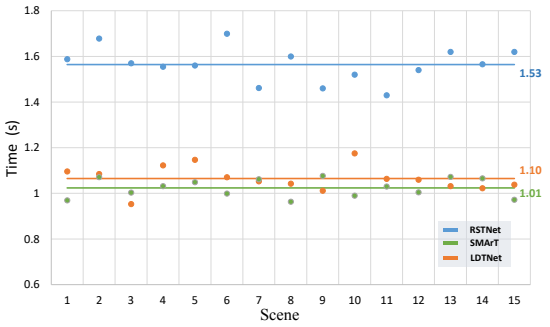


Fig. 7. Response time comparison of image captioning function in our wearable device.

V. CONCLUSIONS

In this paper, we propose a novel lightweight deconfounding network model for image captioning called LDTNet for the first time. On one hand, LDTNet leverages the advantages of shallow features by using inhibitory gating units during the encoding stage, thereby improving the quality of image captions. On the other hand, it has effectively removed confounding factors between the visual and semantics through a causal adjustment module during the decoding stage, thereby significantly enhancing its performance. We also demonstrate the applicability of our method to real-world scene images in actual wearable devices. Extensive experimental results demonstrate its effectiveness. It achieves significant performance improvements over the state of the art and meets the requirements for rapid response. Our future

work plans to adopt recently developed feature selection methods, e.g., [34]–[37], and other deep learning models, e.g., [38]–[42], to further improve the system performance.

REFERENCES

[1] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, 2015, pp. 3156–3164.

[2] X. Zhang, X. Sun, Y. Luo, J. Ji, Y. Zhou, Y. Wu, F. Huang, and R. Ji, "RSTNet: Captioning with adaptive attention on visual and non-visual words," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, 2021, pp. 15 465–15 474.

[3] 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 *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, 2020, pp. 9962–9971.

[4] M. Cornia, L. Baraldi, and R. Cucchiara, "SMArT: Training shallow memory-aware transformers for robotic explainability," in *2020 IEEE Int. Conf. on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 1128–1134.

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

[6] L. Gao *et al.*, "Hierarchical LSTMs with adaptive attention for visual captioning," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 42, no. 5, pp. 1112–1131, 2019.

[7] Y. Pan, T. Yao, Y. Li, and T. Mei, "X-linear attention networks for image captioning," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, 2020, pp. 10 971–10 980.

[8] 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 *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, 2018, pp. 6077–6086.

[9] J. Zhang, Y. Xie, W. Ding, and Z. Wang, "Cross on cross attention: Deep fusion transformer for image captioning," *IEEE Trans. on Circuits and Systems for Video Technology*, 2023.

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[11] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[12] E. Lyu, Z. Zhang, W. Liu, J. Wang, S. Song, and M. Q.-H. Meng, "Mo-transformer: A transformer-based multi-object point cloud reconstruction network," in *2022 IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, 2022, pp. 1024–1030.

[13] C. Yan, Y. Hao, L. Li, J. Yin, A. Liu, Z. Mao, Z. Chen, and X. Gao, "Task-adaptive attention for image captioning," *IEEE Trans. on Circuits and Systems for Video technology*, vol. 32, no. 1, pp. 43–51, 2021.

[14] Y. Luo *et al.*, "Dual-level collaborative transformer for image captioning," in *Proc. of the AAAI Conf. on Artificial Intelligence*, vol. 35, no. 3, 2021, pp. 2286–2293.

[15] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, "Self-critical sequence training for image captioning," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, 2017, pp. 7008–7024.

[16] Y. Li, K. Zhang, J. Cao, R. Timofte, M. Magno, L. Benini, and L. Van Goo, "Localvit: Analyzing locality in vision transformers," in *2023 IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, 2023, pp. 9598–9605.

[17] D. Lopez-Paz *et al.*, "Discovering causal signals in images," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, 2017, pp. 6979–6987.

[18] Z. Yue, H. Zhang, Q. Sun, and X.-S. Hua, "Interventional few-shot learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 2734–2746, 2020.

[19] X. Yang, H. Zhang, and J. Cai, "Deconfounded image captioning: A causal retrospect," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2021.

[20] B. Liu *et al.*, "Show, deconfound and tell: Image captioning with causal inference," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, 2022, pp. 18 041–18 050.

[21] Y. Shi *et al.*, "A dual-feature-based adaptive shared transformer network for image captioning," *IEEE Trans. on Instrumentation and Measurement*, vol. 73, no. 5009613, pp. 1–13, 2024.

[22] K. Xu *et al.*, "Show, attend and tell: Neural image caption generation with visual attention," in *Int. Conf. on Machine Learning*. PMLR, 2015, pp. 2048–2057.

[23] N. Srivastava *et al.*, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[24] L. Huang, W. Wang, J. Chen, and X.-Y. Wei, "Attention on attention for image captioning," in *Proc. of the IEEE/CVF Int. Conf. on Computer Vision*, 2019, pp. 4634–4643.

[25] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, 2015, pp. 4566–4575.

[26] A. Karpathy and F.-F. Li, "Deep visual-semantic alignments for generating image descriptions," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, 2015, pp. 3128–3137.

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

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

[29] S. Banerjee and A. Lavie, "Meteor: An automatic metric for mt evaluation with improved correlation with human judgments," in *Proc. of the Acl Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, 2005, pp. 65–72.

[30] P. Anderson, B. Fernando, M. Johnson, and S. Gould, "Spice: Semantic propositional image caption evaluation," in *Proc. of the European Conf. on Computer Vision (ECCV)*. Springer, 2016, pp. 382–398.

[31] H. Jiang, I. Misra, M. Rohrbach, E. Learned-Miller, and X. Chen, "In defense of grid features for visual question answering," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, 2020, pp. 10 267–10 276.

[32] X. Yang, K. Tang, H. Zhang, and J. Cai, "Auto-encoding scene graphs for image captioning," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, 2019, pp. 10 685–10 694.

[33] M. Cornia, M. Stefanini, L. Baraldi, and R. Cucchiara, "Meshed-memory transformer for image captioning," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, 2020, pp. 10 578–10 587.

[34] J. Luo, Y. Li, Y. Pan, T. Yao, J. Feng, H. Chao, and T. Mei, "Semantic-conditional diffusion networks for image captioning," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, 2023, pp. 23 359–23 368.

[35] Y. Gong *et al.*, "A length-adaptive non-dominated sorting genetic algorithm for bi-objective high-dimensional feature selection," *IEEE/CAA Journal of Automatica Sinica*, vol. 10, no. 9, pp. 1834–1844, Sept. 2023.

[36] Z. Wang, S. Gao, M. Zhou, S. Sato, J. Cheng, and J. Wang, "Information-theory-based nondominated sorting ant colony optimization for multiobjective feature selection in classification," *IEEE Trans. on Cybernetics*, vol. 53, no. 8, pp. 5276–5289, Aug. 2023.

[37] J. Zhou *et al.*, "Lagam: A length-adaptive genetic algorithm with markov blanket for high-dimensional feature selection in classification," *IEEE Trans. on Cybernetics*, vol. 53, no. 11, pp. 6858–6869, Nov. 2023.

[38] H. Zhu *et al.*, "A self-adapting and efficient dandelion algorithm and its application to feature selection for credit card fraud detection," *IEEE/CAA Journal of Automatica Sinica*, vol. 11, no. 2, pp. 377–390, Feb. 2024.

[39] Z. Huang *et al.*, "Feature map distillation of thin nets for low-resolution object recognition," *IEEE Trans. on Image Processing*, vol. 31, pp. 1364–1379, 2022.

[40] Z. Lei *et al.*, "Fully complex-valued gated recurrent neural network for ultrasound imaging," *IEEE Trans. on Neural Networks and Learning Systems*, pp. 1–14, 2023.

[41] P. Xiong *et al.*, "Deeply supervised subspace learning for cross-modal material perception of known and unknown objects," *IEEE Trans. on Industrial Informatics*, vol. 19, no. 2, pp. 2259–2268, Feb. 2023.

[42] Z. Cao *et al.*, "A multi-object tracking algorithm with center-based feature extraction and occlusion handling," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 4, pp. 4464–4473, 2022.