IFNet: An Image-Enhanced Cross-Modal Fusion Network for Radiology Report Generation

Yi Guo, Xiaodi Hou, Zhi Liu, and Yijia Zhang(✉)

School of Information Science and Technology, Dalian Maritime University, Dalian 116026, Liaoning, China
zhangyijia@dlmu.edu.cn

Abstract. Radiology Report Generation (RRG) tasks aim to automatically generate descriptive textual reports for medical images through computer-assisted technologies, which can alleviate the workload of radiologists, reduce the probability of misdiagnoses, and mitigate the strain on medical resources. However, previous studies explore rarely improving low-quality images in datasets, integrating cross-modal information, and optimizing network latency. To address its existing challenges, we propose an **I**mage-enhanced cross-modal **F**usion **Net**work (IFNet) for the automated generation of radiology reports. IFNet comprises three core modules. First is an image enhancement module for augmenting X-ray images' fine-grained normal and abnormal structural representations, increasing the probability of successful detection. Second is a cross-modal fusion network capable of capturing the interactive relations of cross-modal features comprehensively and efficiently. Third is a Transformer report generation module with linear time complexity, aimed at efficiently producing radiology reports with reduced network latency and operability on resource-constrained devices. Experiments on the public dataset IU-Xray demonstrate significant achievements of IFNet, surpassing the performance of the current state-of-the-art methods. The code is available at https://github.com/Hood0602/IFNet.

Keywords: Radiology Report Generation, Medical Image Enhancement, Separable Self-Attention

1 Introduction

The advancement of contemporary medical imaging analysis dramatically promotes medical progress. However, with the increasing number of medical imaging reports, the burden on radiologists escalates, consuming more healthcare resources. Additionally, fatigue and distraction may contribute to radiologists' erroneous judgments. Therefore, the automatic generation of radiology reports has become critical in clinical practice [1, 2, 3, 4, 5].

In recent years, with the rapid development of deep learning and computer vision technology, Radiology Report Generation (RRG) has attracted the attention of researchers and has become a research direction, which can assist the clinical diagnosis and treatment. RRG aims to rapidly and automatically generate medical reports with

lengthy medical descriptions of both standard and pathological areas, assisting doctors in completing diagnostic tasks and alleviating the burden on radiologists.

Existing radiology medical image report generation methods mainly adopt an encoder-decoder framework [1], wherein the encoder extracts global features from radiology datasets while the decoder generates corresponding textual reports. [2] proposed a memory-driven Transformer model to record critical information for report generation. [6] extracts structural information from the corpus to construct a clinical graph based on NLP rules and builds triples for each scenario to replace visual representations with professional knowledge.

Despite some progress made in RRG tasks, several challenges exist:

1. Low-quality dataset: Some images in the X-ray dataset contain severe noise, significantly increasing the difficulty of training and testing the report generation model.
2. Neglect of cross-modal alignment: RRG involves the comprehensive processing of text information and visual information. Most recent studies focus on the processing of single-modal information. This limitation leads to the need for cross-modal information interaction between models.
3. Transformer limitation: Multi-head self-attention mechanisms in Transformers necessitate high computational loads, increasing network latency and posing a significant challenge for resource-constrained devices.

To alleviate the above problems, we propose an **I**mage-Enhanced Cross-Modal **F**usion **Net**work (IFNet), and the specific innovative points are summarized as follows:

1. IFNet efficiently captures the interaction between image and text information through the cross-modal fusion network, thereby improving the accuracy of report generation.
2. We design the image enhancement module *IEC*, which enhances fine-grained features of medical images to improve image quality and information richness, and *SA*Former, which generates reports while reducing latency.
3. To validate the effectiveness of the proposed model IFNet, a series of comprehensive experiments are conducted on the public IU-Xray dataset. The experimental results achieve state-of-the-art performance across various metrics compared to existing baseline models.

2 Methodology

This section explains the specific process by which IFNet obtains cross-modal information and generates radiology reports. Fig. 1. shows the overall structure of IFNet.

2.1 Image Features Extraction

Due to variations in the quality of images generated by different medical imaging devices and the predominant presence of typical structures in these images, the effective identification of abnormal structures becomes exceptionally crucial. Previous me-

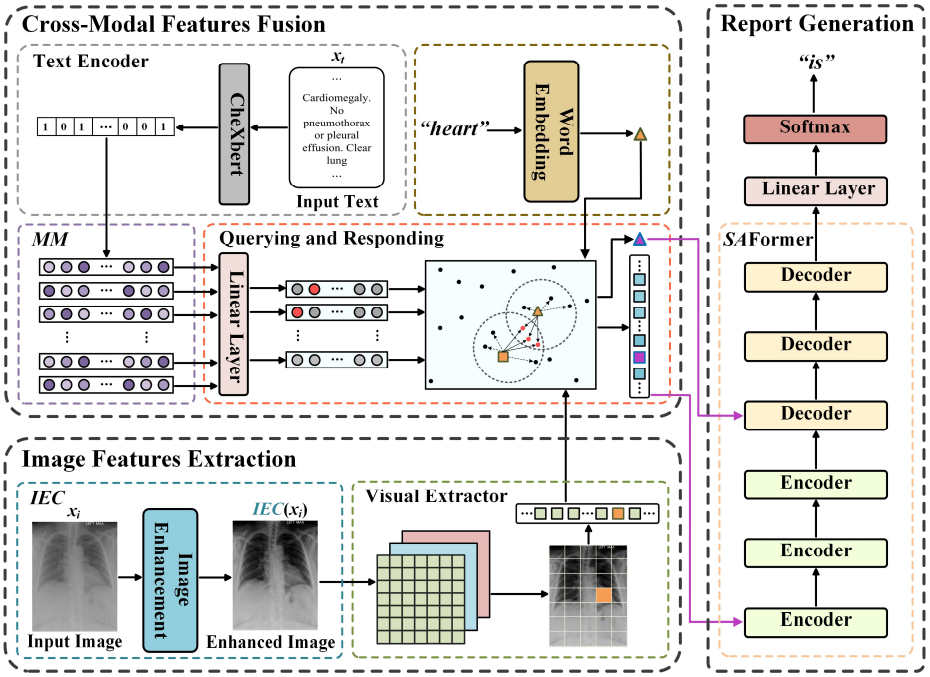

Fig. 1. The overview of the structure of IFNet. IFNet mainly integrates three primary functions: image features extraction, cross-modal features fusion, and report generation.

thods have yet to address maximizing and precisely utilizing the valuable information present in existing datasets while filtering out noise. In response to this challenge, we propose an optimization strategy to enhance the quality of medical images and strengthen the representation of fine-grained abnormal structures.

In IFNet, we introduce an **I**mage **E**nhancement module based on **C**ontrast-limited adaptive histogram equalization (*IEC*), designed to enhance the local contrast of chest X-ray images and thus improve the portrayal of image details in the dataset. Contrast-limited Adaptive Histogram Equalization (CLAHE) adjusts the grayscale distribution of the image in a nonlinear manner, redistributing grayscale values within the target image. Notably, it enhances high-contrast regions of the histogram while attenuating low-contrast areas, achieving a balanced state. This method aims to enhance the visual appearance of images and facilitate the identification of abnormal structures. Specifically, for a given input image x_i, the algorithm of *IEC* creates non-overlapping sub-images, applies histogram equalization to each sub-image, and then clips the original histogram to a specific value before redistributing the clipped pixels to each grayscale level. For the k^{th} image block of the image, we calculate the gain factor for local histogram equalization as follows:

$$G_k(x, y) = OPV(x, y) \cdot CLAHE(x, y) \tag{1}$$

where (x, y) represents the pixel coordinates, $CLAHE(x, y)$ denotes the pixel values

returned using the CLAHE algorithm, and $OPV(x, y)$ represents the original pixel values. Subsequently, we obtain the new pixel values for the k^{th} image block as:

$$NPV_k(x, y) = round(OPV(x, y) \cdot G_k(x, y)) \tag{2}$$

[7] has confirmed the effectiveness of CLAHE in enhancing medical images.

After processing with *IEC*, the image x_i from the dataset yields enhanced image $IEC(x_i)$, which serves as inputs to the CNN. In the feature extraction stage, we utilize a pre-trained ResNet-152 [8] to extract the visual features of $IEC(x_i)$. Specifically, the image features $c \in \mathbb{R}^{H \times W \times C}$ are obtained from the last convolutional layer of the CNN, where H and W represent the height and width of the image, respectively, and C denotes the number of channels. Subsequently, the extracted feature c is linearized, i.e., concatenated row-wise, treating each position's feature as a visual word token. The feature sequence is represented as:

$$\{c_1, c_2, \cdots, c_i, \cdots, c_N\} = f_{ve}(IEC(x_i)) \tag{3}$$

where c_i represents the local feature at the i^{th} position, N is the number of visual word tokens obtained after linearization of feature c, and $f_{ve}(\cdot)$ denotes the visual extractor.

2.2 Cross-Modal Features Fusion

Learning image features and their associated text features is one of the primary challenges in RRG tasks. The cross-modal fusion network allows simultaneous learning of information representations from the image and text and aligning them between different modalities.

After obtaining visual features, we utilize CheXbert [9], a pre-trained model designed for generating labels for chest radiology reports, to acquire the global text features of each report in the dataset. Specifically, by inputting sentences from radiology reports into the model, CheXbert can generate corresponding label values for 14 different observations. The report corresponding to the image x_i is represented as:

$$x_t = \{w_1, w_2, \cdots, w_i, \cdots, w_R\} \tag{4}$$

where w_i represents the i^{th} position's word, and R is the number of words in x_t. The process of labeling can be represented as:

$$\{d_1, d_2, \cdots, d_i, \cdots, d_{14}\} = f_{te}(x_t) \tag{5}$$

where $f_{te}(\cdot)$ denotes the text encoder, $f_{te}(x_t)$ is a one-hot vector, and $d_i \in \{0, 1\}$ is the observation value for one of 14.

The core of our cross-modal fusion network is a memory matrix *MM*, serving as an intermediate state for report generation. We represent *MM* as $\{m_1, m_2, \cdots, m_i, \cdots, m_V\}$ where $MM \in \mathbb{R}^{H \times W \times C}$, $V = N^{te} \times N^{mv}$, N^{te} represents the number of observation results generated by the labeler (CheXbert produces 14 observation results), N^{mv} represents the number of associated categories between x_i and x_t in the dataset, $m_i \in \mathbb{R}^D$ is the memory vector of the i^{th} row, and D is the dimension of the vector.

Initializing the memory matrix with prior information has demonstrated that it can accelerate its convergence process and improve the network's overall performance [10]. We utilize a pre-trained ResNet-152 to extract global visual features, represented as $g^i \in \mathbb{R}^{C_i}$, where C_i is the number of channels extracted from the visual representation. Additionally, CheXbert extracts global text features, representing them as $g^t \in \mathbb{R}^{C_t}$, where C_t is the number of channels extracted from the text representation. Furthermore, to further enhance the model's robustness in handling real-world data, we also extract horizontally and vertically flipped image features $g^{i(\omega)} \in \mathbb{R}^{C_i}$, where $\omega = \{rr, rl, lr, ll\}$ represents the four flip cases of the image. Finally, we obtain a set of feature sets for each observation result the labeler generates, represented as:

$$F_k = \left\{ g_y \middle| d_{v,k} = 1 \right\}, \quad g_y = Concat(g^{i(\omega)}, g^t) \tag{6}$$

Here, F_k represents the cross-modal feature set of category k, and $d_{v,k}$ represents the label of category k for sample v. We obtain the cross-modal feature vector $g \in \mathbb{R}^{(C_i+C_t)}$ by concatenating image and text features.

Subsequently, we utilize the K-Means algorithm [11] to cluster each feature set into N^{mv} clusters, taking the average of the features in each cluster as the initialization value for *MM*. The initialization process is represented as:

$$\{r_1^k, r_2^k, \cdots, r_i^k, \cdots, r_{N^{mv}}^k\} = K\text{-}Means(F_k), \quad r_i^k = \left\{ g_1^{k,i}, g_2^{k,i}, \cdots, g_{N_{k,i}^d}^{k,i} \right\} \tag{7}$$

After obtaining *MM*, inspired by R2GenCMN [3], we employ a multi-threaded querying and responding process to embed the intermediate representation into single-modal features, achieving better interactions between single-modal and cross-modal features. Specifically, given paired inputs $[x_i, x_t]$, the output is the queried cross-modal fusion vector f_v. Before cross-modal querying, to filter out potential noise, f_v is linearly projected to C_F dimensions as follows:

$$f_v = \{MM(k) | d_k = 1\}, \quad f = f_v \cdot W_{fv} \tag{8}$$

where $MM(k)$ represents the vector set of the k^{th} category and $W_{fv} \in \mathbb{R}^{(C_i+C_t)\times C_F}$ is continuously updated during training.

We define that at time point T, the output of the embedding layer for the reports is $\{w_1^e, w_2^e, \cdots, w_i^e, \cdots, w_{T\text{-}1}^e\}$. When querying the vectors to have single-modal features reside in the same feature space, the cross-modal fusion vectors in *MM* and single-modal feature vectors are first linearly transformed as follows:

$$n_i^v = w_i^e \cdot W_s, \quad n_i^t = c_i \cdot W_s, \quad n_i^m = f_i \cdot W_m \tag{9}$$

where $W_s \in \mathbb{R}^{C\times D}$ and $W_m \in \mathbb{R}^{C_F\times D}$ are also continuously updated during training.

As most intermediate state vectors have low relevance to the querying vectors, following [3], the most similar θ vectors (where θ is a hyperparameter) should be selected to eliminate some noise and control memory size. The specific process is as follows: First, we calculate the distance between the visual and textual single-modal features to the corresponding fusion vectors as follows:

$$D^{v}_{(j,\,k)} = \frac{n^{v}_{j} \cdot (n^{m}_{k})^{T}}{\sqrt{D}}, \quad D^{t}_{(j,\,k)} = \frac{n^{t}_{j} \cdot (n^{m}_{k})^{T}}{\sqrt{D}} \tag{10}$$

Then, based on the distance, we determine the similarity between the selected vectors and capture it as follows:

$$\omega^{v}_{(i,\,k)} = \frac{D^{v}_{(i,\,k)}}{\sum_{j=1}^{\theta} D^{v}_{(i,\,j)}}, \quad \omega^{t}_{(i,\,k)} = \frac{D^{t}_{(i,\,k)}}{\sum_{j=1}^{\theta} D^{t}_{(i,\,j)}} \tag{11}$$

After the querying, the querying vectors undergo a linear transformation and weighted sum as follows:

$$\varphi^{v}_{(i,\,j)} = n^{mv}_{(i,\,j)} \cdot W_{m}, \quad \varphi^{t}_{(i,\,j)} = n^{mt}_{(i,\,j)} \cdot W_{m}, \tag{12}$$

$$\gamma^{v}_{i} = \sum_{j=1}^{\theta} \omega^{v}_{(i,\,j)} \cdot \varphi^{v}_{(i,\,j)}, \quad \gamma^{t}_{i} = \sum_{j=1}^{\theta} \omega^{t}_{(i,\,j)} \cdot \varphi^{t}_{(i,\,j)} \tag{13}$$

Here, $n^{mv}_{(i,\,j)}$ and $n^{mt}_{(i,\,j)}$ represent the j^{th} most similar vectors of the i^{th} sub-image and word, respectively. $\varphi^{v}_{(i,\,j)}$ and $\varphi^{t}_{(i,\,j)}$ are the corresponding weighted j^{th} most similar vectors of the i^{th} sub-image and word with θ determining the number of most similar vectors considered for each querying. γ^{v}_{i} and γ^{t}_{i} are the responses of the i^{th} sub-image and word. At this point, the dataset's visual and textual features are loaded into the fusion network to learn the alignment between images and text.

2.3 Report Generation

Many previous works utilize the Multi-Headed Self-Attention (MHA) mechanism to capture the dependencies between positions in sequential data. While this mechanism effectively serves this purpose, it incurs a time complexity of $O(k^2)$ regarding the number of patches k, greatly diminishing the efficiency of the model. Additionally, MHA may entail computationally intensive operations such as matrix multiplication, posing challenges for devices with limited computational capacity, memory, and power constraints. Research addressing network load reduction and optimization efficiency still needs to be completed. Therefore, our work focuses on reducing computational latency and enhancing model efficiency, especially for applications on resource-constrained devices.

In IFNet, we propose a report generation module with ***S**eparable Self-**A**ttention* Trans**former** (*SA*Former). To address the challenges, we introduce *Separable Self-Attention* (SSA) in the encoder and decoder of *SA*Former, as shown in Fig. 2. SSA is a fast and memory-efficient attention mechanism with linear time complexity. Specifically, in SSA, different branches are utilized to process the input $X \in \mathbb{R}^{k \times D}$, namely input I, key K, and value V. Inspired by [12], I is linearly transformed using the learnable weight $W_I \in \mathbb{R}^{D}$ to each memory vector in X, where W_I acts as a latent node L. Subsequently, the distance between L and X is calculated, resulting in a k-dimensional vector, which is then softmax to generate context scores $C_S \in \mathbb{R}^{k}$, as follows:

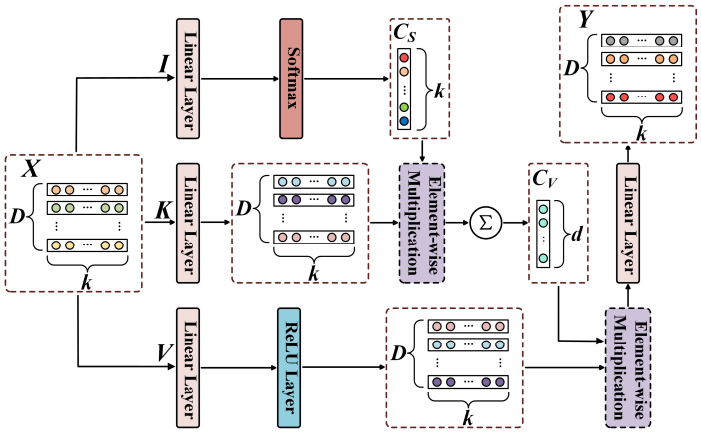

Fig. 2. The internal implementation process of SSA.

$$C_S = Softmax(X \cdot W_I) \tag{14}$$

By not computing context scores for each token relative to all k tokens but only the context scores concerning the latent token L, the time complexity for computing context scores is reduced from $O(k^2)$ to $O(k)$. Subsequently, we utilize the obtained context scores C_S to compute the context vector $C_V \in \mathbb{R}^D$. Specifically, K linearly projects X into D-dimensional space using the learnable weight $W_K \in \mathbb{R}^{D\times D}$, resulting in $X_K \in \mathbb{R}^{D\times D}$. The context vector C_V is then calculated as the weighted sum of X_K, as follows:

$$X_K = X \cdot W_K, \quad C_V = \sum_{i=1}^{k} C_S(i) X_K(i) \tag{15}$$

The integrated contextual information in C_V is shared with the tokens of X. To obtain the final output, V linearly projects X into D-dimensional space using the learnable weight $W_V \in \mathbb{R}^{D\times D}$, followed by ReLU activation to obtain $X_V \in \mathbb{R}^{k\times D}$, expressed as:

$$X_V = ReLU(X \cdot W_V) \tag{16}$$

Subsequently, the integrated contextual information in C_V is propagated to X_V through element-wise multiplication, and the resulting output is input into a linear layer with the learnable weight $W_O \in \mathbb{R}^{D\times D}$ to produce the final output $Y \in \mathbb{R}^{k\times D}$, represented as:

$$Y = (\sum(Softmax(XW_I) * XW_K) * ReLU(XW_V))W_O \tag{17}$$

In the encoder of *SA*Former, the input X for SSA consists of the response-encoded image querying vectors $\{\gamma_1^v, \gamma_2^v, \cdots, \gamma_N^v\}$, as follows:

$$\{E_1^v, E_2^v, \cdots, E_N^v\} = Encoder(\gamma_1^v, \gamma_2^v, \cdots, \gamma_N^v) \tag{18}$$

The input for the decoder is the output of the encoder combined with the textual query vectors for this time point $\{w_1^e, w_2^e, \cdots, w_t^e, \cdots, w_{T\text{-}1}^e\}$, as follows:

$$y_T = Decoder(E_1^v, E_2^v, \cdots, E_N^v, w_1^e, w_2^e, \cdots, w_{T\text{-}1}^e) \tag{19}$$

Repeatedly executing the above process generates the complete radiology report.

3 Experiment Settings

3.1 Dataset and Evaluation Metrics

We validate the effectiveness of IFNet on the public dataset IU-Xray [13], which is widely used in the RRG domain. IU-Xray comprises 7470 X-ray images captured from frontal and lateral views and 3955 medical reports. We employ three standard Natural Language Generation (NLG) metrics (BLEU {1-4} [14], ROUGE-L [15], and METEOR [16]) to assess the model's performance.

3.2 Implementation Details

We use two images of each patient in the IU-Xray dataset and adjust their size to 224 × 224 as input to IFNet. In the visual extractor, we use pre-trained ResNet-152 [8] as the backbone network of CNN to generate each patch feature with 512 dimensions. In the initialization of the cross-modal fusion network, the pre-trained ResNet-152 extracts global visual features with 2048 dimensions, and the pre-trained CheXbert [9] extracts global text features with 768 dimensions. The Clip Limit (CL) used for calculation in *IEC* is set to 1.4, and the sub-image is set to 8 × 8 pixels. The backbone network of the baseline model consists of a Transformer with three layers, eight MHA attention heads, and 512 dimensions, including the encoder and decoder. We use the Adam [17] optimizer and optimize IFNet by cross-entropy loss. The learning rate of the visual extractor is set to $5e-5$, while others are set to $1e-4$.

4 Results and Discussion

4.1 Comparison with Other Methods

To further explore the effectiveness of our approach, we compare it with other methods in the field, including the current state-of-the-art method MMTN [5]. The results are shown in Table 1. Compared to MMTN, IFNet achieves improvements of 2.1%, 2.5%, 1.7%, 1.5%, and 2.8% on NLG metrics BLEU {1-4} and ROUGE-L, respectively. These demonstrate the quality improvement of model-generated reports. The improvement of BLEU demonstrates the high accuracy of words in the model output results. The increase in ROUGE-L demonstrates a high recall rate of the results, more vital information capture ability, and the longest common subsequence (LCS) among longer input sequences. Compared to [4], the METEOR metric significantly improves by 3.4%, indicating that the model can effectively distinguish synonyms and improve sentence fluency. The high performance of IFNet is achieved through three core modules, which we illustrate in subsequent sections.

Table 1. Experimental comparison of IFNet with previous research on NLG metrics based on the IU-Xray dataset. The best values are bolded, and the second-best values are underlined. The metric scores for other methods are quoted from original papers.

MODEL	BL-1	BL-2	BL-3	BL-4	RG-L	MTOR
HRGR [18]	0.438	0.298	0.208	0.151	0.322	-
CoAT [1]	0.455	0.288	0.205	0.154	0.369	-
R2Gen [2]	0.470	0.304	0.219	0.165	0.371	0.187
CMN [3]	0.475	0.309	0.222	0.170	0.375	<u>0.191</u>
PPKED [4]	0.483	0.315	0.224	0.168	<u>0.376</u>	0.190
MMTN [5]	<u>0.486</u>	<u>0.321</u>	<u>0.232</u>	<u>0.175</u>	0.375	-
IFNet (ours)	**0.507**	**0.346**	**0.249**	**0.190**	**0.403**	**0.224**

4.2 Ablation Analysis

We conduct experiments and evaluations on several configurations to investigate the impact of various components in IFNet on its performance. Table 2. shows the results.

Effect of Image Enhancement Module

Based on the baseline model, this configuration integrates the image enhancement module. Compared to the baseline model, NLG metrics BLEU-1, BLEU-2, and BLEU-3 have significantly improved, with values of 5.5%, 3.7%, and 3.3%, as shown in Table 2. The improvement of BLEU {1-3} may be due to *IEC*'s enhancement of fine-grained structure, which increases the probability of detecting normal and abnormal parts of X-ray images. This increases the proportion of vocabulary in the generated report that corresponds to the actual situation.

To maximize the efficiency of *IEC*, we conduct sensitivity experiments on the parameter Clip Limit (CL), as shown in Fig. 3. We use a line chart to visually represent the score of BLEU-1 metric when CL takes various values. When CL increases from 0.9 to 1.4, BLEU-1 score increases and reaches its maximum value (0.507) at 1.4, which is attributed to *IEC*'s practical and appropriate enhancement of image fine-grained structures. As CL continues to increase, BLEU-1 gradually decreases, possibly due to *IEC*'s degree enhancement of images leading to increased noise or loss of details due to overexposure. Therefore, when CL is set to 1.4, it has universality for most images in the dataset, achieving a balance between detail enhancement and overexposure. We show an example of the image obtained under different CL values. As CL value increases, the contrast of the tissues in the image is enhanced. However, when CL value is too high, some noise is also enhanced, and the imaging is distorted.

Effect of Cross-Modal Fusion Network

This setup incorporates the cross-modal fusion network on top of the baseline model. Compared to the baseline, it achieves significant improvements of 7.9%, 6.1%, 5.6%, and 4.3% in NLG metrics BLEU {1-3} and METEOR, respectively, as shown in Tab-

Table 2. Performance evaluation of IFNet components in experiments based on the public dataset IU-Xray. The best values are bolded, and the second-best values are underlined.

MODEL	BL-1	BL-2	BL-3	BL-4	RG-L	MTOR
BASE	0.404	0.260	0.176	0.141	0.347	0.166
BASE + *IEC*	0.459	0.297	0.209	0.162	0.361	0.185
BASE + *MM*	0.483	0.321	0.232	0.161	<u>0.386</u>	0.209
BASE + *IEC* + *MM*	<u>0.504</u>	<u>0.344</u>	<u>0.247</u>	<u>0.183</u>	**0.403**	**0.227**
IFNet	**0.507**	**0.346**	**0.249**	**0.190**	**0.403**	<u>0.224</u>

le 2. The generated reports' word accuracy and sentence fluency can be improved simultaneously by integrating rich cross-modal information. After querying and responding, the interactions between single-modal and cross-modal fusion information in *MM* enable the Transformer to capture additional and accurate cross-modal correspondence, which is the key to this group's far-reaching baseline performance. In addition, *MM* filters out some noise to improve the report's accuracy further.

Effect of *SA*Former

In the two experiments integrating MHA (BASE + *IEC* + *MM*) and SSA (IFNet) in the model, there is no significant difference in metrics, as shown in Table. 2. Still, by calculating only context scores of potential tags L instead of all tags, SSA with linear time complexity can reduce overall network latency while maintaining accuracy. We conduct 30 training epochs on one NVIDIA GeForce RTX 3080 10 GB GPU. On average, *SA*Former reduces network latency to 0.71 times the original value, which means a speed increase of 40.85%. When the batch size value is set to 16, the average peak memory of the MHA group is 8.6 GB, while the SSA group is 5.9 GB. *SA*-Former achieves the same performance with fewer parameters and lower memory consumption, accounting for approximately 68.6% of the original value.

4.3 Case Study

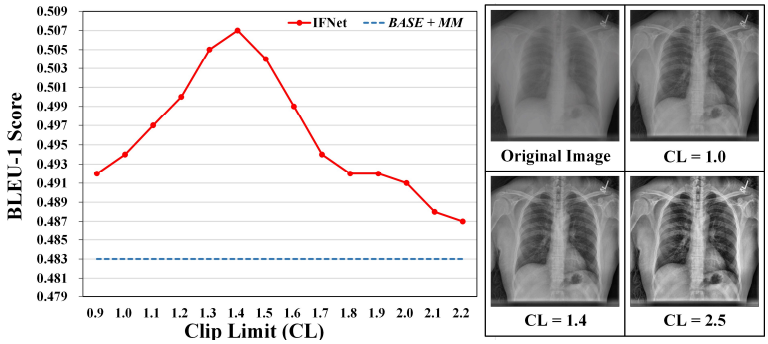

Fig. 3. BLEU-1 scores and enhanced images under different Clip Limit values.

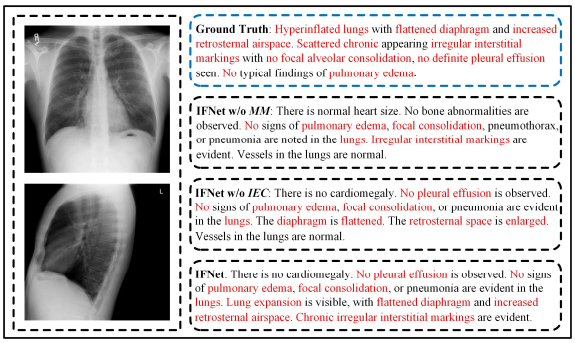

Fig. 4. An illustrative instance of the report visualized with actual data highlighted in red.

Fig. 4. shows a visual example of the reports generated for each configuration. When IFNet removes any component, the accuracy of generating reports, the number of words containing the truth decreases, and the proportion of invalid report words increases. Specifically, when removing *MM*, the model's ability to capture relationships decreases, making it unable to capture complex relationships, e.g., "*Hyperinflated lungs with flattered diaphragm*". When removing *IEC*, capturing details e.g. as "*interstitial markings*" decreases. When *MM* combines cross-modal information with *IEC*'s enhancement of details, reports generated by IFNet are the closest to the ground truth.

5 Limitations

Although IFNet achieves excellent performance, there is a gap between the generated report and the description provided by doctors, and the accuracy still needs to reach a sufficient level for clinical application. Any missed or misdiagnosis caused by data deviation is unacceptable. In future work, we aim to integrate warning mechanisms to prompt doctors' intervention when difficult-to-identify lesions occur.

6 Conclusions

This paper introduces an **I**mage-Enhanced Cross-Modal **F**usion **Net**work (IFNet) designed to enhance the detail of image structures, fortify the fusion of single-modal and cross-modal interaction information, and reduce network and hardware resource consumption while automating the generation of radiology reports. We develop three core modules to achieve these objectives. Experimental findings on the openly accessible IU-Xray dataset validate the effectiveness of IFNet. We also conduct ablation analysis to demonstrate the effectiveness of the components within the model for performance enhancement. IFNet surpasses current state-of-the-art methods, with all three core modules significantly improving the model's performance.

References

1. Jing, B., Xie, P., & Xing, E. (2017). On the automatic generation of medical imaging reports. *arXiv preprint arXiv:1711.08195*.
2. Chen, Z., Song, Y., Chang, T. H., & Wan, X. (2020). Generating radiology reports via memory-driven transformer. *arXiv preprint arXiv:2010.16056*.
3. Chen, Z., Shen, Y., Song, Y., & Wan, X. (2022). Cross-modal memory networks for radiology report generation. *arXiv preprint arXiv:2204.13258*.
4. Liu, F., Wu, X., Ge, S., Fan, W., & Zou, Y. (2021). Exploring and distilling posterior and prior knowledge for radiology report generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 13753-13762).
5. Cao, Y., Cui, L., Zhang, L., Yu, F., Li, Z., & Xu, Y. (2023, June). MMTN: multi-modal memory transformer network for image-report consistent medical report generation. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 37, No. 1, pp. 277-285).
6. Li, M., Cai, W., Verspoor, K., Pan, S., Liang, X., & Chang, X. (2022). Cross-modal clinical graph transformer for ophthalmic report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 20656-20665).
7. Sahu, S., Singh, A. K., Ghrera, S. P., & Elhoseny, M. (2019). An approach for de-noising and contrast enhancement of retinal fundus image using CLAHE. *Optics & Laser Technology*, *110*, 87-98.
8. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).
9. Smit, A., Jain, S., Rajpurkar, P., Pareek, A., Ng, A. Y., & Lungren, M. P. (2020). CheXbert: combining automatic labelers and expert annotations for accurate radiology report labeling using BERT. *arXiv preprint arXiv:2004.09167*.
10. Wang, J., Bhalerao, A., & He, Y. (2022, October). Cross-modal prototype driven network for radiology report generation. In *European Conference on Computer Vision* (pp. 563-579). Cham: Springer Nature Switzerland.
11. Lloyd, S. (1982). Least squares quantization in PCM. *IEEE transactions on information theory*, *28*(2), 129-137.
12. Mehta, S., & Rastegari, M. (2022). Separable self-attention for mobile vision transformers. *arXiv preprint arXiv:2206.02680*.
13. Demner-Fushman, D., Kohli, M. D., Rosenman, M. B., Shooshan, S. E., Rodriguez, L., Antani, S., ... & McDonald, C. J. (2016). Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, *23*(2), 304-310.
14. Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002, July). 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).
15. Lin, C. Y. (2004, July). Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out* (pp. 74-81).
16. Denkowski, M., & Lavie, A. (2011, July). Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In *Proceedings of the sixth workshop on statistical machine translation* (pp. 85-91).
17. Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
18. Li, Y., Liang, X., Hu, Z., & Xing, E. P. (2018). Hybrid retrieval-generation reinforced agent for medical image report generation. *Advances in neural information processing systems*, *31*.