TSGET: Two-Stage Global Enhanced Transformer for Automatic Radiology Report Generation

Xiulong Yi, You Fu, Ruiqing Liu, Hao Zhang, and Rong Hua

***Abstract*—Recently, automatic radiology report generation, which targets to generate multiple sentences that can accurately describe medical observations for given X-ray images, has gained increasing attention. Existing methods commonly employ the attention mechanism for accurate word generation. However, such attention-based methods fail to leverage useful image-level global features, thereby limiting the model's reasoning ability. To tackle this challenge, we propose two-stage global enhancement layers to facilitate the Transformer to generate more reliable reports from a global perspective. Specifically, the 1st Global Enhancement Layer (1st GEL) is designed to capture the global visual context features by establishing the relationships between image-level global features and previously generated words. The 2nd Global Enhancement Layer (2nd GEL) is devised to capture the region-global level features by building the relationships between image-level global features and region-level information. The experiments demonstrate that by integrating the aforementioned two-stage global enhancement layers into the Transformer model, our proposal achieves state-of-the-art (SOTA) performance on various Natural Language Generation (NLG) evaluation metrics. Further Clinical Efficacy (CE) evaluations also validate that our proposal is able to predict more critical information.**

***Index Terms*—Transformer, radiology report generation, attention mechanism, global enhancement.**

I. Introduction

ANALYZING radiology images and preparing corresponding reports are indispensable steps for radiologists in the current clinical diagnosis process. However, the analysis and interpretation of radiology images is a time-consuming and arduous task for experienced radiologists, leading to inefficient utilization of medical resources. Inexperienced radiologists often face significant challenges in producing error-free reports, resulting in the potential oversight of some abnormalities. In this situation, the automation of generating accurate radiology reports is of great clinical value, which has gained increasing attention [1], [2], [3].

Transformer: The lungs are clear without focal consolidation. No pleural effusion or pneumothorax is seen . The cardiac and mediastinal silhouettes are unremarkable .

Ours: The lungs are clear . There is no pleural effusion or pneumothorax . The heart size is normal . The mediastinal and hilar contours are normal . Rib fractures is seen.

Fig. 1. Example of report generated by the Transformer and our method.

Recently, attention mechanism is prevalent in current automatic radiology report generation methods. For example, Jing et al. [4] design a multi-task learning framework, where a Multi-Label Classification process is first adopted to extract semantic features, and then a Co-Attention module is devised to combine the extracted semantic features and the original visual features. On this basis, Park et al. [5] take full advantage of the feature differences between the patient and normal images to facilitate the model to focus more on abnormal findings. Liu et al. [6] propose the Contrastive Attention to distill the contrastive features, which can better solve the serious data bias in automatic radiology report generation task. Shi et al. [7] propose the Merging Gate to flexibly combine the extracted semantic features and the grid features in an adaptive manner.

Even though the aforementioned attention-based methods have achieved remarkable performance, they still fail to leverage valuable image-level global features, resulting in object omissions during report generation. As Fig. 1 shows, the Transformer model ignores an important concept, "Rib fractures". In addition, attention-based methods treat each representations in isolation and ignore global guidance, which will lead to a relation bias in different visual representations.

The key to solving the above problems is to capture and utilize the global features to guide the decoding process. In this paper, we directly employ the mean pooling operation for extracting

Manuscript received 2 May 2023; revised 6 December 2023; accepted 31 December 2023. Date of publication 5 January 2024; date of current version 5 April 2024. This work was supported by the National Natural Science Foundation of China under Grant 82000482. *(Corresponding author: Rong Hua.)*

Xiulong Yi, You Fu, and Rong Hua are with the College of Computer Science and Engineering, Shandong University of Science and Technology, Qingdao, Shandong 266590, China (e-mail: yxl1620159241@163.com; fuyou@sdust.edu.cn; huarong@sdust.edu.cn).

Ruiqing Liu is with the Department of Gastrointestinal Surgery, The Affiliated Hospital of Qingdao University, Qingdao, Shandong 266003, China (e-mail: liuruiqing@qdu.edu.cn).

Hao Zhang is with the Department of Computer Science, University of Otago, Dunedin 9054, New Zealand (e-mail: hao.zhang@postgrad.otago.ac.nz).

The Code will be available at. https://github.com/SKD-HPC/TSGET.

Digital Object Identifier 10.1109/JBHI.2024.3350077

global features from given X-ray images, and the primary focus of our work lies in the effective utilization of global features. Specifically, based on the Transformer [8], we propose the TSGET model, which integrate the global features with other multi-modal features to facilitate the base model to generate more reliable reports from a global perspective. In TSGET, we first design the 1st global enhancement layer to capture the global visual context features. To this end, we adopt a single Linear layer followed by a ReLU activation function to fuse the image-level global features with previously generated words. The output of our 1st global enhancement layer is fed to the Masked Multi-Head Attention to generate current hidden state. Then we design the 2nd global enhancement layer to capture the local-global level features. In detail, we first construct the Refined Module, which can determine whether or how well both the global feature and attended feature are correlated with the current hidden state, thereby facilitating the elimination of redundant feature components. Subsequently, we adopt the Fusion Module to fuse the refined global feature and the refined attended feature in a adaptive way.

We evaluate the impact of applying both the 1st and 2nd global enhancement layers to the Transformer [8] respectively, and the ablation results demonstrate its effectiveness. Besides, comparison results on two benchmark datasets show that by integrating the above two-stage global enhancement layers into the Transformer model, our proposal achieves SOTA performance on various NLG evaluation metrics. Further CE evaluation also validate that our proposal is able to predict more accurate medical information.

Our contributions are summarized as follows:

- We propose the TSGET model, which contains two-stage global enhancement layers, to alleviate the problem of object missing and relation bias occurred in popular attention-based automatic radiology report generation methods.
- We propose two-stage global enhancement layers to introduce the image-level global features into the decoding process, where the 1st GEL is responsible for capturing the global visual context features, and the 2nd GEL is responsible for capturing the region-global level features.
- Experiments on two datasets illustrate that by integrating the above two-stage global enhancement layers into the Transformer model, our proposal achieves SOTA performance on various NLG evaluation metrics.

II. Related Works

A. Image Captioning

Image captioning, which aims to generate a concise sentence describing an image, is the most prevalent task related to automated radiology report generation. Existing image captioning methods [9], [10], [11] are mostly based on the attention mechanism, which can ground correct objects and obtain attended feature for proper word generation. For example, Lu et al. [12] propose the Adaptive Attention, which can adaptively decide whether to rely on images or language models. Ji et al. [13] propose the Spatio-Temporal Memory Attention, which can extend the traditional attention mechanism to learn the spatio-temporal relationships. Ke et al. [14] propose the Dual Attention to exploit the cooperation between textual and visual. In addition, X-Linear Attention [15] is devised to exploit the interaction between intra- and inter-modal features. Attention on Attention [16] is devised to solve the irrelevant attention problem in the generation process. In recent years, various attention-based methods are proposed for image captioning. Most of them are fit for automatic radiology report generation. However, instead of generating a short sentence, automatic radiology report generation methods usually need to generate long paragraphs that can accurately describe the observations for given X-ray images. Therefore, Multi-Head Attention based Transformer model [8], which has a better ability to handle long-term dependencies, is more popular than conventional attention mechanisms in automatic radiology report generation task.

B. Transformer-Based Radiology Report Generation

Recently, Transformer is widely used in current automatic radiology report generation methods [17], [18]. [19]. For example, the R2Gen model [20] proposes Memory-Driven Transformer to take full advantage of similar patterns in different X-ray images. The pure-Transformer model [21] adopts the pre-trained vision transformer [22] as the visual extractor instead of traditional CNN-based methods. The GSKET model [23], which can introduce the General and Specific Knowledge into the decoding process to generate more reliable reports. In addition, Cross-modal Memory Network [24] is designed to build interaction across modalities. On this basis, Reinforced Cross-modal Memory Network [25] employs reinforcement learning to provide appropriate supervision from NLG evaluation metrics. Progressive Transformer [26] divides the full generation process into three stages. The Prior Guided Transformer [27] is proposed to introduce prior knowledge into the decoding process.

The above Transformer-based automatic radiology report generation methods achieve remarkable performance, which can better build the interaction between grid visual features and textual features by the cross-attention module. However, such methods neglect the effect of image-level global information, which causes object missing and relation bias problem when generating reports, and limits the reasoning capability of the model to some extent.

C. Model With Global Features

There are several works aiming at exploring the effect of global features in different image-to-text tasks. For example, Zhang et al. [28] propose a GVFGA and a LSGA module for remote sensing image captioning. The formal is proposed to introduce global features into the decoding process and remove redundant components. The latter is proposed to reduce the hidden state's burden. Ji et al. [29] proposes the Global-Enhanced Transformer for image captioning, where a Global Enhanced Encoder is adopted to extract global features, and a Global Adaptive Decoder is adopted to introduce the global features into the decoding process to guide the caption generation. Li

TABLE I
NOTATIONS USED IN THIS PAPER

Notations	Descriptions
f_v	The visual extractor.
n	The number of extracted grid features.
I,y	Input images, and ground-truth reports.
$\overline{v}$	The image-level global features.
f_e, f_d	The Transformer Encoder and Decoder.
θ	The model parameters.
$\widehat{v}$	The attended feature
T	The length of ground-truth reports.
$r(\cdot)$	The reward function in RL.
f_g	The image-level global feature extractor.
σ	The sigmoid activation function.
H	The hidden state.
W, b	The learnable weights.
f_r	The Refined Module.
f_s	The score function.
C_t	The context features.
l_s	The length of ground-truth sentences.
P, R	The accuracy and the recall.
l_c	The length of generated sentences.
$\overline{v}'$, $\widehat{v}'$	The refined global feature and refined attended feature.
LCS	The length of the longest common sub-sequence.

et al. [30] proposes the Global-Local Attention for image captioning, which can combines the image-level and region-level information.

The aforementioned works illustrate that incorporating global features can enhance the performance of various image-to-text tasks, resulting in the generation of captions with superior quality. Therefore, in automatic radiology report generation task, Yin et al. [31] propose the global label pooling mechanism for predicting abnormalities directly from extracted features. Liu et al. [32] introduce the global pooling to generate the hidden state of LSTM. Different from the previous works [31], [32], our proposal introduces global features into the decoding process by fusing the global features with other multi-modal features (attended feature and textual feature).

III. METHOD

We describe our TSGET model in this section, which integrates the proposed two-stage global enhancement layers into the Transformer model [8] for automatic radiology report generation. Table I shows the major notations and their corresponding descriptions used in this paper.

A. Overview

As Fig. 2 shows, TSGET model includes three major components:

Visual Extractor Following most radiology report generation methods [20] [24], the original ResNet-101 [33], pretrained on ImageNet, is employed to extract the grid features from the given X-ray images:

$$\{S_1, \ldots, S_n\} = f_m(f_v(I)) \tag{1}$$

where f_v refers to the visual extractor, f_m refers to a single-layer linear function, I refers to the input X-ray images, and n refers to the number of extracted grid feature.

Encoder Given the extracted grid features $\{S_1, \ldots, S_n\}$, we adopt the Transformer encoder to further build the interaction between different grid features:

$$\{V_1, \ldots, V_n\} = f_e(S_1, \ldots, S_n) \tag{2}$$

where f_e represents the Transformer encoder.

Decoder At every time step t, given the encoded grid features V and the target words y, we adopt the decoder to generate current word y_t. To facilitate the Transformer to generate more reliable reports from a global perspective, we integrate the proposed two-stage GELs into the Transformer decoder for word generation:

$$y_t = f_d((1^{st}(\overline{v}, y), 2^{nd}(\overline{v}, H, \widehat{v})) \tag{3}$$

where f_d refers to the decoder, $\widehat{v}$ refers to the attended feature (generated by (7)), $\overline{v}$ refers to the global feature (generated by (8)), H refers to the hidden state (generated by (11)).

B. Cross-Attention

We first provide a description of traditional Multi-Head Attention, which is formulated as follows:

$$MHA(Q, K, V) = [H_1, H_2, \ldots, H_h]W^o \tag{4}$$

$$H_i = Att\left(QW_i^Q, KW_i^K, VW_i^V\right) \tag{5}$$

$$Att(Q, K, V) = \text{softmax}\left(\frac{QK^{\mathrm{T}}}{\sqrt{d_k}}\right)V \tag{6}$$

where W^o, W_i^Q, W_i^K, W_i^V are learnable parameters, [,] refers to the concatenation operation, H_i refers to the ith head in Multi-Head Attention, h refers to the number of heads, and $\sqrt{d_k}$ refers to the scaling factor.

The Cross-Attention module, based on the Multi-Head Attention mechanism, serves as an intermediary that establishes a connection between the encoder and decoder, thereby facilitating the processing of multi-modal features (visual and textual). In Cross-Attention, the hidden state usually serves as the query vector to generate the attended feature from encoded grid features:

$$\widehat{v} = MHA(H, V, V) \tag{7}$$

where H refers to the current hidden state, and $V = \{V_1, \ldots, V_n\}$ refers to the encoded grid features.

C. The 1st Global Enhancement Layer

In Transformer, hidden state is usually used as the query vector to generate the attended feature from encoded grid features. As shown in Fig. 3(a), most existing Transformer-based automatic radiology report generation methods generate the current hidden state by directly model the relationships between generated words, where a Masked Multi-Head Attention is used to achieve this purpose.The calculation process of the hidden state, however, fails to consider the impact of the global feature that represents the entire image. As Fig. 3(b) shows, we propose the 1st GEL to provide global visual context features for the Masked Multi-Head Attention. Details of the devised 1st GEL are described in Fig. 2.

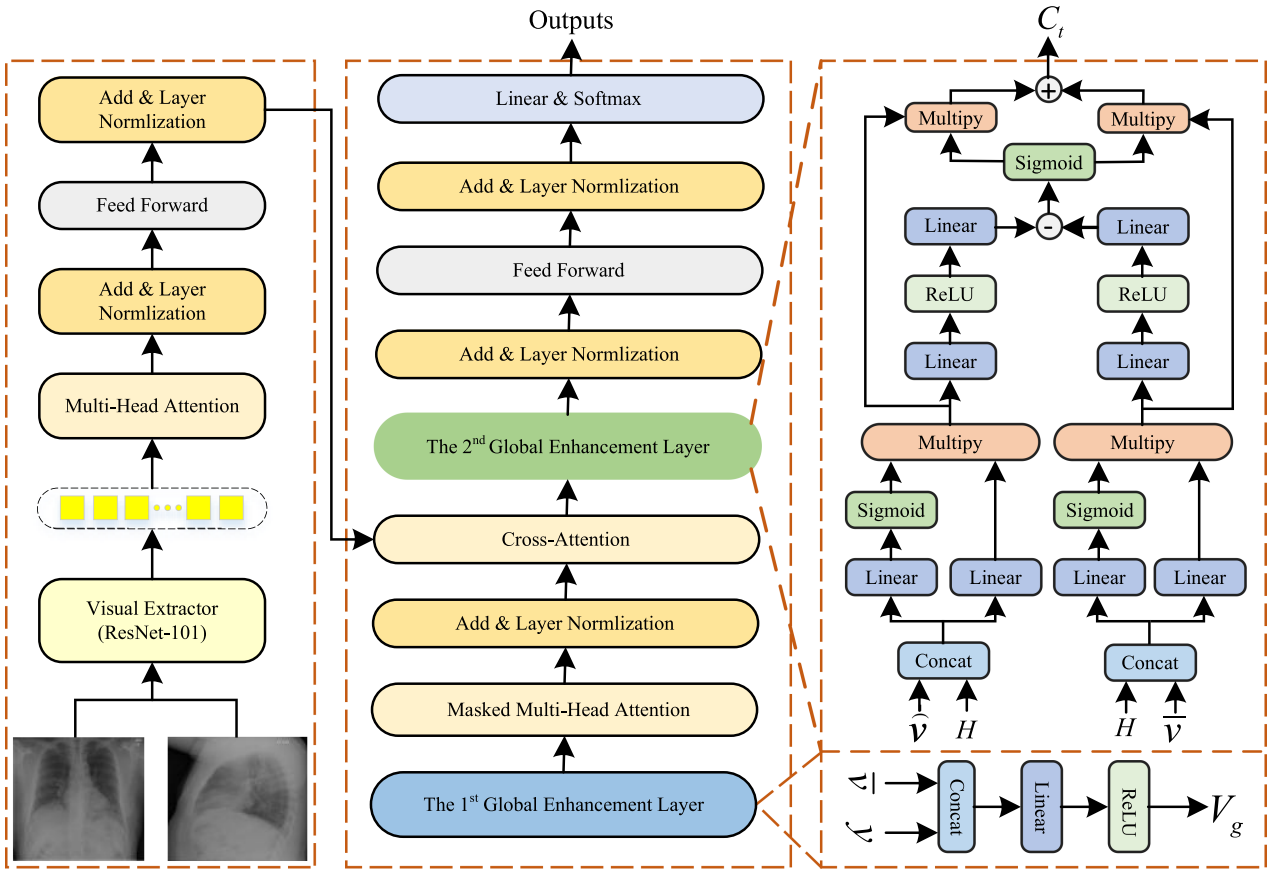

Fig. 2. Illustration of our proposed TSGET model, which follows the conventional encoder-decoder framework. For encoder, a pre-trained ResNet is utilized to extract original grid features from input images, followed by a three-layer Transformer encoder that generate the encoded grid features. For decoder, we integrate the proposed 1st and 2nd GELs into Transformer decoder to facilitate the Transformer to generate more reliable reports from a global perspective.

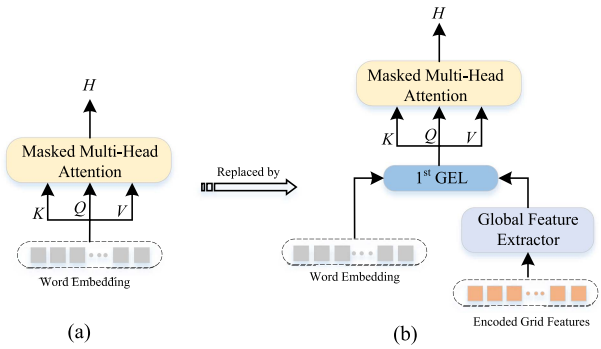

Fig. 3. The calculation process of the hidden state. (a) Traditional methods; (b) Ours, where the 1st GEL is proposed to provide global visual context features.

To this end, we first adopt a Global Feature Extractor to generate the global feature:

$$\overline{v} = f_g(V_1, \ldots, V_n) \tag{8}$$

where $\{V_1, \ldots, V_n\}$ are encoded grid features, and f_g refers to the Global Feature Extractor. Inspired by [30], in this paper, we taking the mean pooling operation as the Global Feature Extractor.

Then, we adopt a word embedding to generate the representations of target words:

$$Y = W_e y \tag{9}$$

where W_e is the word embedding, and y refers to the target words.

Finally, at every time step t, given the global representation $\overline{v}$, and the target word representations Y, we design the 1st GEL to provide global visual context features for the Masked Multi-Head Attention, which is formulated as follows:

$$V_g = ReLU(W_g[Y, \overline{v}] + b_g) \tag{10}$$

where W_g, b_g are learnable weights to fuse the global feature with the target word representations, and [,] refers to concatenation.

The output of the 1st GEL V_g is fed to the Masked Multi-Head Attention to generate current hidden state H:

$$H = f_m(V_g, V_g, V_g) \tag{11}$$

where f_m refers to the Masked Multi-Head Attention.

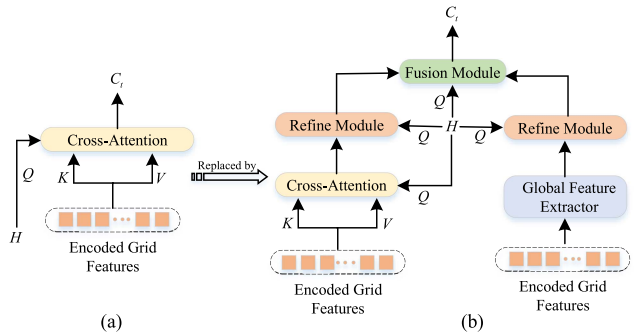

Fig. 4. The calculation process of the context features. (a) Traditional methods; (b) Ours, Where the 2nd GEL is proposed to capture the region-global level information.

D. The 2nd Global Enhancement Layer

As shown in Fig. 4(a), in Transformer-based automatic radiology report generation methods, Cross-Attention is used to build the cross-modal interaction between encoded grid features and previously generated words, whose output is usually used to predict the next word. However, such methods suffer from the following problems when directly used for automatic radiology report generation tasks. 1) The Cross-Attention invariably returns a weighted average vector regardless of whether candidate features are meet the requirement of query vector, which will returns a fallacious result in some cases. 2) The Cross-Attention only builds the interaction between encoded grid features and previously generated words, which neglects the effect of image-level global information.

To solve the above problems, as shown in Fig. 4(b), after performing Cross-Attention, we construct the 2nd global enhancement layer to capture the attended-global level information. Details of the proposed 2nd GEL are illustrated in Fig. 2. In detail, the proposed 2nd global enhancement layer contains two major components: the Refined Module, and the Fusion Module.

Refined Module We first construct the Refined Module for removing irrelevant information, which is defined as follows:

$$f_r(q,x) = \sigma(W_g^q q + W_g^x x + b_g) \odot W_i^q q + W_i^x x + b_i \quad (12)$$

where W_g^q, W_g^x, b_g, W_i^q, and b_i are learnable weights, σ refers to the sigmoid activation function, $\odot$ refers to element-wise multiplication, and q refers to the query vector.

Then, given the attended feature $\widehat{v}$ (generated by (7)) and global feature $\overline{v}$ (generated by (8)), we take hidden state H (generated by (11)) as the query vector to generate the refined attended feature and the refined global feature by:

$$\widehat{v}' = f_r(H, \widehat{v}) \quad (13)$$

$$\overline{v}' = f_r(H, \overline{v}) \quad (14)$$

Fusion Module Given the output of Refined Module, we construct the Fusion Module to fuse the refined global feature $\overline{v}'$ and the refined attended feature $\widehat{v}'$, which can help to obtain more comprehensive visual information of input images. To this end, we first design a score function to calculate how much attention should be given to refined attended feature and global feature, respectively. The score function is defined as follows:

$$f_s(u) = W_s(RELU(W_s^u)u + b_s) \quad (15)$$

where W_s, b_s are learnable weights. Then, the context features C_t can be calculated by the following formula:

$$\alpha_t = \sigma(f_s(\widehat{v}') - f_s(\overline{v}')) \quad (16)$$

$$C_t = \alpha_t \widehat{v}' + (1-\alpha)\overline{v}' \quad (17)$$

where $\alpha_t \in [0,1]$ refers to the importance of refined attended feature $\widehat{v}'$ compared to the refined global feature $\overline{v}'$. The Fusion Module makes the model can adaptively focus on global feature or refined attended feature.

E. Training

The entire training process consists of two stages: initially pretraining the TSGET model using a word-level Cross-Entropy loss, followed by fine-tuning the sequence generation through Reinforcement Learning. When training with Cross-Entropy loss, the model is trained to predict the next word given previous target words:

$$L = -\sum_{t=1}^{T} \log(p_\theta(y_t | y_{1:t-1})) \quad (18)$$

where $y_{1:t-1}$ refers to the target words, θ refers to the model parameters, and T represents the length of target reports. After several epoch Cross-Entropy pretraining stage, we start to optimize the proposed TSGET model with Reinforcement Learning. In implementation, our TSGET model is perceived as an **agent** that interacts with textual and visual features (**environment**). The parameters of TSGET model θ define a **policy** P_θ that results in an **action**. Upon generating the end-of-sequence token, the agent observes a **reward** r. The reward is computed by BLEU metric, which is formulated as follows:

$$BLEU = \min\left(1, \exp\left(1 - \frac{l_s}{l_c}\right) \times \exp\left(\sum_{n=1}^{N} \omega_n \log P_n(C,S)\right)\right) \quad (19)$$

$$P_n(C,S) = \frac{\sum_i \sum_k \min(h_k(c_i), \max_{j\in m} h_k S(i,j))}{\sum_i \sum_k \min(h_k(c_i))} \quad (20)$$

where N usually is 1, 2, 3, and 4; $\omega_n = 1/N$ is a hyper-parameter; C refers to the generated text, l_c refers to its length, and $c_i \in C$ refers to the ith generated word; Similarly, S refers to the target text, l_s refers to its length, and $s_i \in S$ refers to the ith target word; h_k refers to the number of times n-grams of length k occur in the text. Therefore, the loss can be defined as follows:

$$L(\theta) = -E_{y\sim P_\theta}[r(Y)] \quad (21)$$

where Y refers to the generated words.

IV. Experiments and Analysis

A. Datasets

We conduct experiments on the following two benchmark datasets:

- Indiana University Chest X-ray Collection (IU X-Ray[1]) [34] was released in 2015 by Indiana University, which includes 7,470 images and 3,995 reports. However, IU-Xray dataset does not have a standardized data split method. In this paper, to ensure fair comparison with previous studies, we adopt the dataset split provided by [20], which randomly divides the dataset into training, validation, and testing sets at a ratio of 7:1:2 without any overlap in patients.
- MIMIC-CXR[2] dataset [35] was released in 2019 by Beth Israel Deaconess Medical Center, which includes 473,057 images and 236,563 corresponding reports. The official split method is adopted in this paper for MIMIC-CXR.

B. Evaluation Metrics

We first adopt the traditional NLG evaluation metrics[3] to verify the effectiveness of our TSGET model. The NLG evaluation metrics include: BLEU [36], METEOR [37], and ROUGE-L [38]. It is worth mentioning that the evaluation scores of conventional NLG metrics, such as BLEU-n and ROUGE, solely rely on a limited subset of words. Consequently, these metrics may not offer comprehensive assessments in terms of medical diagnosis. In this paper, we adopt CE metrics to analyze the quality of generated reports from the perspective of clinics. To this end, we adopt the CheXpert[4] to label the generated reports, and utilize the Precision, Recall, and F1-Score to evaluate the ability to describe the abnormalities.

C. Implementation Details

For the visual extractor, we first adopt the pre-trained ResNet-101 [33] to extract grid features with a dimension of 7*7*2,048, and then a single fully connected layer is used to convert the dimension of grid features to 7*7*512. For the TSGET model, the number of heads is set to 8, the dimension of Multi-Head Attention is set to 512, and the number of encoder-decoder layers is set to 3. For training, on IU X-Ray dataset [34], we first train the TSGET model for 100 epochs under Cross-Entropy loss with the mini-batch 16, and then followed by Reinforcement Learning for 80 epochs with the mini-batch 10. We use both the frontal and lateral images as input, and the max length of generated reports is set to 60. On the MIMIC-CXR dataset [35], we first train the TSGET model for 40 epochs under the cross-entropy loss with the mini-batch 16 and then followed by Reinforcement Learning for 12 epochs with the mini-batch 6. We use a single image as input, and the max length of generated reports is set to 100. For both datasets, the learning rate is set to 5e-5. All experiments are performed under the following experimental conditions. CPU: E5-2690 V4, GPU: NVIDIA Tesla P100, CUDA version: 10.0, python version: 3.6, pytorch version: 1.7.0.

[1]https://openi.nlm.nih.gov/
[2]https://physionet.org/content/mimic-cxr/2.0.0/
[3]https://github.com/tylin/coco-caption
[4]https://github.com/MIT-LCP/mimic-cxr/tree/master/txt/chexpert

D. Comparison With Previous Methods

1) Comparison With Previous Methods on NLG Metrics: We compare the proposed TSGET model with the following SOTA methods on NLG Metrics.

1) AdaAtt [12]: Adaptive Attention, which can adaptively decide whether to rely on images or language model.
2) M2Transformer [39]: Meshed Memory Transformer employs a novel meshed connectivity between the encoder and decoder.
3) Grounded [40]: It is designed to generate sentences by distilling the image-textual matching information.
4) Co-Att [4]: A Hierarchical LSTM based method, which can localize abnormalities by the proposed Co-Attention mechanism.
5) CMAS-RL [47]: It is designed to exploit the structure information in different reports.
6) R2Gen [20]: It is designed to generate long sentences by distilling the similar patterns in different reports.
7) CMCL [41]: Competence-based Multimodal Curriculum Learning for remedying the data bias problem, which can easily integrated to the Hierarchical LSTM based methods.
8) CMN [24]: It is designed to exploit the cross-modal mappings between images and texts.
9) PGT [27]: Prior Guided Transformer, which is designed to exploit the influence of prior knowledge.
10) PPKED [42]: It is designed to exploit the posterior and prior Knowledge.
11) Align-Transformer [43]: A Transformer-based method, which can align the grid features and the disease tag features
12) M2TR [26]: A novel framework, which divides the generation process into two steps (image-text-text).
13) CA [6]: Contrastive Attention, which is designed to accurately describe the abnormal regions.
14) PureT [21]: The first pure Transformer-based method, which extracts visual features without using CNN.
15) GSKET [23]: It is designed to exploit both the general and specific knowledge.

Comparison results are presented in Table II . For image captioning methods: AdaAtt [12], M2Transformer [39], and Grounded [40], we quote the results from published literature [21]. As for other methods, we refer to the original papers for their reported results.

There are several observations obtained from Table II. Firstly, in comparison with LSTM-based image captioning methods AdaAtt [12], Grounded [40], and Transformer-based image captioning method M2Transformer [39], our proposal outperforms all of them on all NLG metrics. This observation highlights the necessity of designing a specific module for automatic radiology report generation task. Secondly, our proposal outperforms the

TABLE II
COMPARISON OF EVALUATION SCORES BETWEEN OUR METHOD AND OTHER METHODS

Dataset	Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	Rouge-L
IU X-Ray [34]	AdaAtt [12]	43.6	28.8	20.3	15.0	-	35.4
	M2Transformer [39]	46.3	31.8	21.4	15.5	-	33.5
	Grounded [40]	44.6	30.1	23.7	17.6	-	34.3
	Co-Att [4]	45.5	28.8	20.5	15.4	-	36.9
	CMAS-RL [47]	46.4	30.1	21.0	15.4	-	36.2
	R2Gen [20]	47.0	30.4	21.9	16.5	18.7	37.1
	CMCL [41]	47.3	30.5	21.7	16.2	18.6	37.8
	CMN [24]	47.5	30.9	22.2	17.0	19.1	37.5
	PGT [27]	48.2	31.3	23.2	18.1	20.3	38.1
	PPKED [42]	48.3	31.5	22.4	16.8	-	37.6
	Align-Transformer [43]	48.4	31.3	22.5	17.3	20.4	37.9
	M2TR [26]	48.6	31.7	23.2	17.3	19.2	39.0
	CA [6]	49.2	31.4	22.2	16.9	19.3	38.1
	PureT [21]	49.6	31.9	24.1	17.5	-	37.7
	GSKET [23]	49.6	32.7	23.8	17.8	-	38.1
	TSGET (Ours)	**50.0**	**34.9**	**25.6**	**19.4**	**21.8**	**40.2**
MIMIC-CXR [35]	AdaAtt [12]	29.9	18.5	12.4	8.8	-	26.6
	M2Transformer [39]	21.2	12.8	8.3	5.8	-	24.0
	Grounded [40]	27.1	17.4	12.2	9.4	-	25.7
	R2Gen [20]	35.3	21.8	14.5	10.3	14.2	27.7
	CMCL [41]	34.4	21.7	14.0	9.7	13.3	28.1
	CMN [24]	35.3	21.8	14.8	10.6	14.2	27.8
	PGT [27]	35.6	22.2	15.1	11.1	14.0	28.0
	PPKED [42]	36.0	22.4	14.9	10.6	14.9	28.4
	Align-Transformer [43]	37.8	23.5	15.6	11.2	**15.8**	28.3
	M2TR [26]	37.8	23.2	15.4	10.7	14.5	27.2
	CA [6]	35.0	21.9	15.2	10.9	15.1	28.3
	PureT [21]	35.1	22.3	15.7	11.8	-	**28.7**
	GSKET [23]	36.3	22.8	15.6	11.5	-	28.4
	TSGET (Ours)	**39.8**	**24.8**	**16.9**	**12.1**	14.9	28.1

Results are illustrated as percentage (%).
The best results are highlighted in bold.

Hierarchical LSTM-based automatic radiology report generation methods Co-Att [4], and CMAS-RL [47] by a significant margin. This observation indicates that Transformer [8] better suited for automatic radiology report generation tasks due to its superior ability to handle long-term dependencies. Thirdly, Our proposal exhibits a significant performance improvement compared to other Transformer-based automatic radiology report generation methods, R2Gen [20], CMN [24], M2TR [26], PGT [27], Align-Transformer [43], and PureT [21]. Specifically, on the IU X-Ray dataset and the MIMIC-CXR dataset, our approach achieves an increase in BLEU-4 scores from 18.1 to 19.4 and from 11.8 to 12.1 respectively. This observation highlights the efficacy of fusing global features with multi-modal features for generating high-quality reports. Fourthly, compared to some complicated methods CMCL [41], PPKED [42], CA [6], and GSKET [23], our proposal is simpler and more effective.

2) Comparison With Previous Studies on CE Metrics: We further compare the TSGET model with several SOTA methods on CE Metrics. The compared methods include TopDown [52], M2TR [26], PGT [27], CA [6], R2Gen [20], CMN [24], RL-CMN [25], CMCA [53], and GSKET [23]. As Table III shows, our proposed TSGET model significantly outperforms previous methods with a Recall increase from 23.1 to 50.9 and an F1-Score increase from 23.8 to 39.3. These comparison results demonstrate the effectiveness of our approach.

TABLE III
PERFORMANCE COMPARISON OF DIFFERENT METHODS ON CE METRICS

Dataset	Method	Precision	Recall	F1-Score
MIMIC-CXR [35]	TopDown [52]	32.0	23.1	23.8
	M2TR [26]	24.0	42.8	30.8
	PGT [27]	35.3	31.0	29.7
	CA [6]	35.2	29.8	30.3
	R2Gen [20]	33.3	27.3	27.6
	CMN [24]	33.4	27.5	27.8
	RL-CMN [25]	34.2	29.4	29.2
	CMCA [53]	44.4	29.7	35.6
	GSKET [23]	**45.8**	34.8	37.1
	TSGET(Ours)	31.9	**50.9**	**39.3**

Results are illustrated as percentage (%).
The best results are highlighted in bold.

E. Ablation Study

Our base model is the Transformer [8] without any additional modifications. To evaluate the contributions of each proposed component, we conducted ablation experiments on both IU X-Ray and MIMIC-CXR datasets. Experimental results are presented in Table IV, where "Avg" denotes the average performance improvement of our proposed TSGET model over the Base model across all NLG metrics. Moreover, "Ours-XE" and "Ours-RL" respectively refer to the TSGET models trained using Cross-Entropy and Reinforcement Learning techniques. The proposed TSGET model achieves an average performance improvement of 24.0% and 20.3% on two benchmark datasets, respectively, as demonstrated in Table IV. This observation serves to underscore the effectiveness of our proposal. Additionally, we present visualizations of the performance of all ablation methods (trained solely under Cross-Entropy Loss) on three evaluation metrics in Figs. 5, 6, and 7. The visualization results

TABLE IV
RESULTS OF ABLATION STUDIES CONDUCTED ON IU X-RAY AND MIMIC-CXR DATASETS, WHICH ARE ILLUSTRATED AS PERCENTAGE (%)

Dataset	Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	Rouge-L	Avg
IU X-Ray [34]	Base	42.4	27.2	19.6	15.1	17.6	35.1	-
	Base+1^{st} GEL	47.7	31.0	22.7	17.6	19.1	36.3	11.8%
	Base+2^{nd} GEL	47.5	32.2	24.2	18.9	19.7	39.4	17.2%
	Ours-XE	49.6	33.5	24.9	19.2	20.0	39.3	20.0%
	Ours-RL	**50.0**	**34.9**	**25.6**	**19.4**	**21.8**	**40.2**	**24.0%**
MIMIC-CXR [35]	Base	32.3	19.9	13.3	9.5	12.8	27.2	-
	Base+1^{st} GEL	34.3	21.0	14.1	10.1	13.7	27.5	5.4%
	Base+2^{nd} GEL	34.9	21.5	14.4	10.2	13.9	27.6	7.0%
	Ours-XE	35.6	22.2	15.0	10.8	14.1	27.9	10.2%
	Ours-RL	**39.8**	**24.8**	**16.9**	**12.1**	**14.9**	**28.1**	**20.3%**

The best results are highlighted in bold.

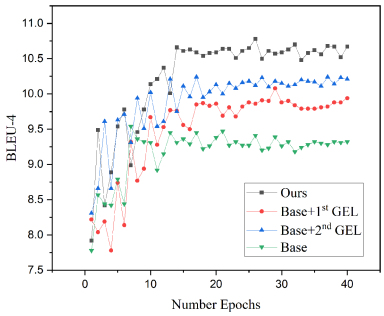

Fig. 5. Performance evaluation of all ablation methods on BLEU-4 metric [36].

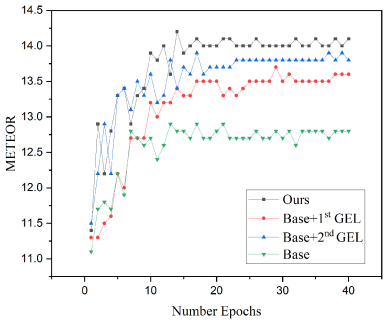

Fig. 6. Performance evaluation of all ablation methods on METEOR metric [37].

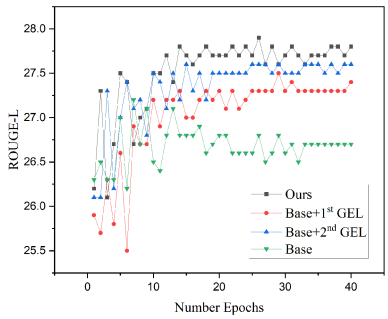

Fig. 7. Performance evaluation of all ablation methods on ROUGE-L metric [38].

provide a more intuitive demonstration of the effectiveness of each proposed component.

1) Effect of the 1st GEL: To assess the efficacy of the 1st GEL, before performing Masked Multi-Head Attention, we employ it to capture the global visual context features. The implementation details are illustrated in Fig. 3(b). As shown in Table IV, the Base + 1st GEL achieves 11.8% and 5.4% average performance improvement on two benchmark datasets, respectively. This underscores the critical role played by the global visual context features extracted by the 1st GEL in generating more dependable reports.

2) Effect of the 2nd GEL: To evaluate the effectiveness of the 2nd GEL, we employ it to capture region-global level information after performing Cross-Attention. The implemented details are shown in Fig. 4(b). As shown in Table IV, the Base + 2nd GEL yields an average performance improvement of 17.2% and 7.0% on two benchmark datasets, respectively. This suggests that extracting region-global level information with the aid of the 2nd GEL can significantly enhance the base model's performance.

3) Effect of the RL: To assess the efficacy of RL, we employ it to provide appropriate supervision from NLG metrics after training multiple epochs under Cross-Entropy. As shown in Table IV, Ours-RL model achieves an average performance improvement of 4.0% and 10.1% on two benchmark datasets (IU X-Ray and MIMIC-CXR), respectively, compared to Ours-XE model. Notably, the most significant enhancement is observed on the MIMIC-CXR dataset, where RL significantly boosts the performance of our Ours-XE model.

F. Discussion

In this section, we first conduct experiments to investigate the impact of different visual feature extractors. Subsequently, we examine the effect of various hyper-parameters on model performance, including beam size and number of Transformer layers. Finally, we discuss the complexity and general applicability of our proposed TSGET model. All experiments in this section are trained using Cross-Entropy loss exclusively.

1) Influence of Various Visual Extractors: We conduct experiments to investigate the influence of various visual extractors, including traditional CNN-based methods: VggNet-16 [44], VggNet-19 [44], GoogLeNet [48], ResNet-18 [33], ResNet-50 [33], ResNet-101 [33], ResNet-152 [33], DenseNet-121 [45], DenseNet-169 [45] as well as recent recent Transformer-based

TABLE V
PERFORMANCE COMPARISON OF THE TSGET MODEL UNDER VARIOUS VISUAL EXTRACTORS

Dataset	Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	Rouge-L
IU X-Ray [34]	VggNet-16 [44]	46.2	29.5	21.6	16.9	18.4	35.2
	VggNet-19 [44]	47.1	30.8	22.7	17.7	19.5	36.5
	GoogLeNet [48]	42.3	27.2	19.9	15.5	17.5	35.5
	ResNet-18 [33]	42.1	27.9	21.0	16.9	17.8	35.7
	ResNet-50 [33]	43.1	28.1	20.9	16.5	17.6	35.7
	ResNet-101 [33]	**49.6**	**33.5**	**24.9**	**19.2**	**20.0**	**39.3**
	ResNet-152 [33]	43.8	28.5	21.3	16.9	18.0	35.9
	DenseNet-121 [45]	44.2	29.8	22.1	16.9	18.8	39.0
	DenseNet-169 [45]	41.1	27.0	20.2	16.2	17.3	35.7
	Swin-Transformer-B [46]	46.0	28.9	21.0	16.3	18.3	35.1
	Swin-Transformer-L [46]	47.6	30.8	22.4	17.3	19.7	36.6
MIMIC-CXR [35]	VggNet-19 [44]	32.4	19.9	13.3	9.4	13.0	27.2
	GoogLeNet [48]	34.6	21.5	14.5	10.4	13.9	**28.0**
	ResNet-50 [33]	34.5	21.3	14.2	10.1	13.7	27.5
	ResNet-101 [33]	**35.6**	**22.2**	**15.0**	**10.8**	**14.1**	27.9
	DenseNet-121 [45]	34.9	21.8	14.9	**10.8**	**14.1**	**28.0**
	Swin-Transformer-L [46]	30.9	19.2	13.1	9.5	12.7	26.9

Results are illustrated as percentage (%).
The best results are highlighted in bold.

TABLE VI
PERFORMANCE COMPARISON OF THE TSGET MODEL UNDER VARYING NUMBERS OF TRANSFORMER LAYERS ON THE IU X-RAY DATASET

Dataset	Layers	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	Rouge-L
IU X-Ray [34]	1	41.8	27.6	20.4	16.2	17.8	36.4
	2	45.5	28.9	21.0	16.4	18.2	35.7
	3	**49.6**	**33.5**	**24.9**	**19.2**	**20.0**	**39.3**
	4	44.3	29.1	21.6	17.1	18.2	36.4
	5	46.3	30.3	22.4	17.5	18.8	36.7
	6	44.2	28.3	20.8	16.4	17.9	34.8

Results are illustrated as percentage (%).
The best results are highlighted in bold.

TABLE VII
COMPARISON OF MODEL PARAMETERS, PERFORMANCE, AND TRAINING EFFICIENCY AMONG THREE METHODS IS CONDUCTED ON THE IU X-RAY DATASET

Dataset	Method	Parameters	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	Rouge-L	Time
IU X-Ray [34]	Base	56.95 M	42.4	27.1	19.6	15.1	17.6	35.1	65s
	R2Gen [20]	**78.47 M**	47.0	30.4	21.9	16.5	18.7	37.1	**106s**
	Ours	66.41 M	**49.6**	**33.5**	**24.9**	**19.2**	**20.0**	**39.3**	70s

Results are illustrated as percentage (%).
The best results are highlighted in bold.

methods: Swin-Transformer-B [46], and Swin-Transformer-L [46]. As demonstrated in Table V, we found that ResNet-101 is the most suitable visual feature extractor for our TSGET model.

2) Influence of the Transformer Layers: In this study, we conduct experiments to investigate the impact of varying numbers of Transformer layers on model performance. As shown in Table VI, our results indicate that increasing the number of Transformer layers initially improves evaluation scores before reaching a point where further increases lead to diminishing returns. For consistency across all experiments, we set the number of Transformer layers at 3.

3) Complexity Analysis: We conduct experiments to analyze the complexity of the proposed TSGET model, To this end, comparing it with the R2Gen [20] and and Base models in terms of model parameters, performance, and training efficiency. Table VII presents the experimental results, revealing two key observations. Firstly, the proposed TSGET model outperforms the recent SOTA R2Gen model [20] across all evaluation metrics while utilizing fewer model parameters. Secondly, training one epoch with the proposed TSGET model takes approximately 70 seconds, which is 36 seconds faster than the R2Gen model and only five seconds slower than the base model. To conclude, the aforementioned observations demonstrate that our proposal achieves superior performance at a minimal cost to training efficiency.

4) Generalization Analysis: We conduct experiments to investigate the general applicability of the proposed TSGET model. To this end, we integrate the proposed 1st and 2nd global enhancement layers into an existing SOTA method: S2-Transformer [49]. The S2-Transformer model is proposed for image captioning, where a Spatial-aware Pseudo-supervised module is designed for preserving spatial information, and a Scale-wise Reinforcement module is designed for exploiting hierarchical encoded features. In our experiments, we follow the original settings in [49], where the weighting factor $\lambda = 0.2$, and the number of clusters $N = 5$. As Table VIII shows, the proposed two-stage global enhancement layers exhibit a significant improvement in the performance of S2-Transformer, resulting in an average performance gain of 3.63% and 4.15% on two benchmark datasets respectively. This observation demonstrates

TABLE VIII
GENERAL APPLICABILITY ANALYSIS RESULTS ON TWO BENCHMARK DATASETS

Dataset	Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	Rouge-L	Avg
IU X-Ray [34]	S2-Transformer [49]	43.8	28.9	21.1	16.3	18.4	36.3	-
	S2-Transformer+Ours	**45.6**	**30.0**	**22.2**	**17.4**	**18.5**	**36.8**	**3.63%**
MIMIC-CXR [35]	S2-Transformer [49]	33.2	20.6	13.9	10.0	13.4	27.6	-
	S2-Transformer+Ours	**35.0**	**21.7**	**14.6**	**10.5**	**13.9**	**27.7**	**4.15%**

Results are illustrated as percentage (%).
The best results are highlighted in bold.

TABLE IX
PERFORMANCE COMPARISON OF THE TSGET MODEL UNDER VARIOUS BEAM SIZES

Dataset	Beam	BLEU-1	BLEU-2	BLEU-3	BLEU-4
IU X-Ray [34]	1	43.8	27.4	19.0	14.1
	2	49.1	33.3	24.8	19.1
	3	**49.6**	**33.5**	**24.9**	**19.2**
	4	43.5	28.4	21.4	17.2
	5	43.2	28.3	21.3	17.1

Results are illustrated as percentage (%).
The best results are highlighted in bold.

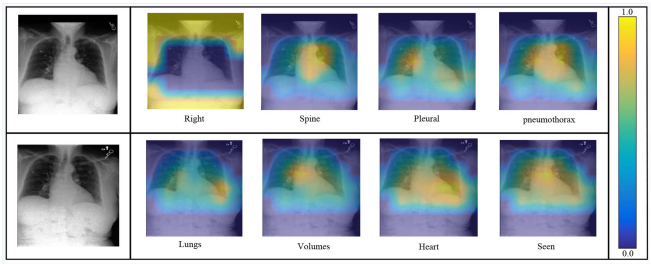

Fig. 8. Visualization of image-text attention mappings during the report generation process of the proposed TSGET model, where the colors ranging from blue to yellow indicate the corresponding attention weights from low to high.

the general applicability of the proposed 1st and 2nd global enhancement layers, which also hold great potential for other Transformer-based methods. Note that the performance of the S2-Transformer + Ours is inferior to that of the TSGET model, possibly due to the utilization of visual cluster information and image-level global representation leading to redundant information.

5) Influence of the Beam Size: Beam Search is an improvement over the conventional greedy strategy, as it no longer selects words solely based on their score but instead retains a wider search of potential candidates. As shown in Table IX, we conducted experiments to verify the impact of different beam sizes and found that increasing the size initially improves evaluation metrics before eventually leading to diminishing returns. The beam size is thus set to 3.

6) Attention Visualization: We conduct experiments to demonstrate the image-text attention mappings of our proposed TSGET model. As Fig. 8 shows, the TSGET model effectively establishes interactions between images and generated words, such as "spine", and "heart". These observations indicate that by integrating two-stage global enhancement layers into the Transformer model, our proposed TSGET model can provide more precise visual information for the decoder.

V. CONCLUSION

We propose the TSGET model as a solution to alleviate the problems of object omission and relation bias that are prevalent in current attention-based automatic radiology report generation methods. The TSGET model comprises two proposed layers, where the 1st GEL is responsible for integrating image-level global features with previously generated words to capture the global visual context features, and the 2nd GEL is responsible for fusing image-level global features with region-level information to obtain more comprehensive visual information. The integration of image-level global features with these multimodal features facilitates the model to generate more reliable reports from a global perspective. Experimental results demonstrate that our method, which incorporates the aforementioned two-stage global enhancement layers into the Transformer model, achieves SOTA performance on various NLG and CE metrics.

Although the proposed TSGET model has achieved satisfactory performance, there is still room for improvement. Firstly, the serious data bias problem poses a challenge to the proposed TSGET model. Exploring the impact of disease tags may offer a potential solution to this issue. Secondly, while using ImageNet [50] as pre-training for visual extractor in this paper is effective, fine-tuning ResNet-101 on CheXpert dataset [51] may further enhance our model's performance.

REFERENCES

[1] C. Y. Li, X. Liang, Z. Hu, and E. P. Xing, "Hybrid retrieval-generation reinforced agent for medical image report generation," in *Proc. 32nd Int. Conf. Neural Inf. Process. Syst.*, 2018, pp. 1537–1547.
[2] C. Y. Li, X. Liang, Z. Hu, and E. P. Xing, "Knowledge-driven encode, retrieve, paraphrase for medical image report generation," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 6666–6673.
[3] Y. Zhang, X. Wang, Z. Xu, Q. Yu, A. Yuille, and D. Xu, "When radiology report generation meets knowledge graph," in *Proc. AAAI Conf. Artif. Intell.*, vol. 2020, pp. 12910–12917.
[4] B. Jing, P. Xie, and E. Xing, "On the automatic generation of medical imaging reports," in *Process. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 2577–2586.
[5] H. Park, K. Kim, J. Yoon, S. Park, and J. Choi, "Feature difference makes sense: A medical image captioning model exploiting feature difference and tag information," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics: Student Res. Workshop*, vol. 2020, pp. 95–102.
[6] X. Ma et al., "Contrastive attention for automatic chest X-ray report generation," in *Proc. Findings Assoc. Computat. Linguistics*, 2021, pp. 269–280.
[7] J. Shi, S. Wang, R. Wang, and S. Ma, "AIMNet: Adaptive image-tag merging network for automatic medical report generation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2022, pp. 7737–7741.
[8] A. Vaswani et al., "Attention is all you need," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 6000–6010.
[9] F. Sammani and L. Melas-Kyriazi, "Show, Edit and tell: A framework for editing image captions," in *Proc. IEEE/CVF Conf. Computer Vis. Pattern Recognit.*, 2020, pp. 4808–4816.

[10] W. Jiang, M. Zhu, Y. Fang, G. Shi, X. Zhao, and Y. Liu, "Visual cluster grounding for image captioning," *IEEE Trans. Image Process.*, vol. 31, pp. 3920–3934, 2022.

[11] L. Wang, Z. Bai, Y. Zhang, and H. Lu, "Show, recall, and tell: Image captioning with recall mechanism," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 12176–12183.

[12] J. Lu, C. Xiong, D. Parikh, and R. Socher, "Knowing when to look: Adaptive attention via a visual sentinel for image," in *Proc. IEEE Conf. Computer Vis. Pattern Recognit.*, 2017, pp. 375–383.

[13] J. Ji, C. Xu, X. Zhang, B. Wang, and X. Song, "Spatio-temporal memory attention for image captioning," *IEEE Trans. Image Processing.*, vol. 29, pp. 7615–7628, 2020.

[14] L. Ke, W. Pei, R. Li, X. Shen, and Y.-W. Tai, "Reflective decoding network for image," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, vol. 2019, pp. 8888–8897.

[15] Y. Pan, T. Yao, Y. Li, and T. Mei, "X-linear attention networks for image," in *Proc. IEEE/CVF Conf. Computer Vis. Pattern Recognit.*, 2020, pp. 10971–10980.

[16] L. Huang, W. Wang, J. Chen, and X. -Y. Wei, "Attention on attention for image," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 4634–4643.

[17] Z. Wang, L. Liu, L. Wang, and L. Zhou, "METransformer: Radiology report generation by transformer with multiple learnable expert tokens," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, vol. 2023, pp. 11558–11567.

[18] Z. Huang, X. Zhang, and S. Zhang, "KiUT: Knowledge-injected U-transformer for radiology report generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 19809–19818.

[19] J. Zhang et al., "A Novel deep learning model for medical report generation by inter-intra information calibration," *IEEE J. Biomed. Health Inform.*, vol. 27, no. 10, pp. 5110–5121, Oct. 2023.

[20] Z. Chen, Y. Song, T.-H. Chang, and X. Wan, "Generating radiology reports via memory-driven transformer," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2020, pp. 1439–1449.

[21] Z. Wang, H. Han, L. Wang, X. Li, and L. Zhou, "Automated radiographic report generation purely on transformer: A. multicriteria supervised approach," *IEEE Trans. Med. Imag.*, vol. 41, no. 10, pp. 2803–2813, Oct. 2022.

[22] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Representations*, 2021.

[23] S. Yang, X. Wu, S. Ge, S. K. Zhou, and L. Xiao, "Knowledge matters: Chest radiology report generation with general and specific knowledge," *Med. Image Anal.*, vol. 80, 2022, Art. no. 102510.

[24] Z. Chen, Y. Shen, Y. Song, and X. Wan, "Cross-modal memory networks for radiology report generation," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics, 11th Int. Joint Conf. Natural Lang. Process.*, 2021, pp. 5904–5914.

[25] H. Qin and Y. Song, "Reinforced cross-modal alignment for radiology report generation," in *Proc. Findings Assoc. Computat. Linguistics*, 2022, pp. 448–458.

[26] F. Nooralahzadeh, N. P. Gonzalez, T. Frauenfelder, K. Fujimoto, and M. Krauthammer, "Progressive transformer-based generation of radiology reports," in *Proc. Findings Assoc. Comput. Linguistics: EMNLP*, 2021, pp. 2824–2832.

[27] B. Yan, M. Pei, M. Zhao, C. Shan, and Z. Tian, "Prior guided transformer for accurate radiology reports generation," *IEEE J. Biomed. Health Inform.*, vol. 26, no. 11, pp. 5631–5640, Nov. 2022.

[28] Z. Zhang, W. Zhang, M. Yan, X. Gao, K. Fu, and X. Sun, "Global visual feature and linguistic state guided attention for remote sensing image captioning," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–16, 2022.

[29] J. Ji et al., "Improving image captioning by leveraging intra-and inter-layer global representation in transformer network," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 1655–1663.

[30] L. Li, S. Tang, L. Deng, Y. Zhang, and Q. Tian, "Image caption with global-local attention," in *Proc. 31st AAAI Conf. Artif. Intell.*, 2017, pp. 4133–4139.

[31] C. Yin et al., "Automatic generation of medical imaging diagnostic report with hierarchical recurrent neural network," in *Proc. IEEE Int. Conf. Data Mining*, 2019, pp. 728–737.

[32] G. Liu et al., "Clinically accurate chest X-ray report generation," in *Proc. 4th Mach. Learn. Healthcare Conf.*, 2019, pp. 249–269.

[33] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[34] D. Demner-Fushman et al., "Preparing a collection of radiology examinations for distribution and retrieval," *J. Amer. Med. Inform. Assoc.*, vol. 23, no. 2, pp. 304–310, 2016.

[35] A. E. W. Johnson et al., "MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports," *Sci. Data*, vol. 6, no. 1, p. 317, Dec. 2019.

[36] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics*, 2002, pp. 311–318.

[37] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in *Proc. ACL Workshop Intrinsic Extrinsic Eval. Meas. Mach. Transl. Summarization*, 2005, pp. 65–72.

[38] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Proc. Workshop Text Summarization Branches Out*, 2004, pp. 74–81.

[39] M. Cornia, M. Stefanini, L. Baraldi, and R. Cucchiara, "Meshed-memory transformer for image," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 10578–10587.

[40] Y. Zhou, M. Wang, D. Liu, Z. Hu, and H. Zhang, "More grounded image captioning by distilling image-text matching model," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, vol. 2020, pp. 4777–4786.

[41] F. Liu, S. Ge, and X. Wu, "Competence-based multimodal curriculum learning for medical report generation," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics 11th Int. Joint Conf. Natural Lang. Process.*, 2021, pp. 3001–3012.

[42] F. Liu, X. Wu, S. Ge, W. Fan, and Y. Zou, "Exploring and distilling posterior and prior knowledge for radiology report generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, vol. 2021, pp. 13753–13762.

[43] D. You et al., "Aligntransformer: Hierarchical alignment of visual regions and disease tags for medical report generation," in *Proc. Med. Image Comput. Computer Assist. Interv.*, vol. 2021, pp. 72–82.

[44] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Rep.*, 2015, pp. 1–10.

[45] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Computer Vis. Pattern Recognit.*, 2017, pp. 4700–4708.

[46] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, vol. 2021, pp. 10012–10022.

[47] B. Jing, Z. Wang, and E. Xing, "Show, describe and conclude: On exploiting the structure information of chest X-ray reports," in *Proc. Assoc. Comput. Linguistics*, 2019, pp. 6570–6580.

[48] C. Szegedy et al., "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, vol. 2015, pp. 1–9.

[49] P. Zeng, H. Zhang, J. Song, and L. Gao, "S2 transformer for image captioning," in *Proc. 31st Int. Joint Conf. Art. Intell.*, 2022, pp. 1608–1614.

[50] J. Deng, W. Dong, R. Socher, L. -J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.

[51] J. Irvin et al., "CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 590–597.

[52] P. Anderson et al., "Bottom-up and top-down attention for image captioning and visual question answering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6077–6086.

[53] X. Song, X. Zhang, J. Ji, Y. Liu, and P. Wei, "Cross-modal contrastive attention model for medical report generation," in *Proc. 29th Int. Conf. Comput. Linguistics*, 2022, pp. 2388–2397.