Bridging the Gap: Cross-modal Knowledge Driven Network for Radiology Report Generation

Beichen Kang
bckang21@m.fudan.edu.cn
Shanghai Key Laboratory of Data
Science, School of Computer Science,
Fudan University, Shanghai, China

Yun Xiong*
yunx@fudan.edu.cn
Shanghai Key Laboratory of Data
Science, School of Computer Science,
Fudan University, Shanghai, China

Jianbo Jiao
j.jiao@bham.ac.uk
School of Computer Science,
University of Birmingham,
Birmingham, United Kingdom

Yao Zhang
yaozhang@fudan.edu.cn
Shanghai Key Laboratory of Data
Science, School of Computer Science,
Fudan University, Shanghai, China

Xing Jia
xjia18@fudan.edu.cn
Shanghai Key Laboratory of Data
Science, School of Computer Science,
Fudan University, Shanghai, China

Ji Li
liji@huashan.org.cn
Department of Pancreatic Surgery,
Huashan Hospital,
Fudan University, Shanghai, China

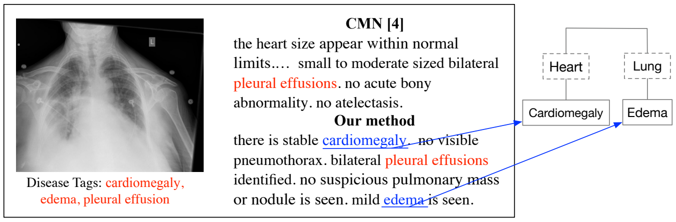
***Abstract*—Radiology report generation aims to generate medical reports based on given medical images, which can alleviate the workload of radiologists and has attracted significant research interest in recent years. However, existing studies have struggled to bridge the gap between the two different modalities (i.e. image and text) and generate clinically accurate reports. This is primarily due to the challenges in modelling the cross-modal mappings and the inefficiency of transferring knowledge across modalities. To address these challenges, in this paper, we propose to leverage a pre-constructed knowledge graph as a shared matrix that bridges the gap between visual and textual information, facilitating cross-modal knowledge transfer. This shared knowledge matrix effectively captures cross-modal mappings and aligns information between images and texts, thereby bridging the gap between modalities. Specifically, we propose a new module for knowledge distillation and preservation that integrates relevant knowledge representations into both visual and textual inputs, facilitating intuitive cross-modal knowledge interaction and enhancing the clinical accuracy of the generated reports. Experimental results on two benchmark datasets show the effectiveness of our method, outperforming state-of-the-arts in report generation.**

***Index Terms*—Radiology Report Generation, Multimodal, Graph, Medical Data Mining**

I. Introduction

Radiology report writing and medical image interpretation (e.g., chest X-rays) are essential in clinical practice and often involve a substantial manual workload. Thus, there is a strong desire for radiology report generation, which automates the generation of textual descriptions using radiology images, to alleviate the heavy workload of radiologists while ensuring healthcare quality.

Automatic radiology report generation has attracted significant research attention in recent years [1]–[8]. Radiology report generation is a cross-modal task, with the majority of existing methods relying on the standard image captioning paradigm [9], [10], which employs a conventional encoder-decoder architecture. Although these methods have achieved remarkable performance, they primarily focus on text generation side and do not fully exploit the information across radiology images and reports. As a result, the findings may remain unsatisfactory, posing challenges in identifying certain diseases and resulting in low clinical accuracy in the generated reports.



Disease Tags: cardiomegaly, edema, pleural effusion

CMN [4]

the heart size appear within normal limits.... small to moderate sized bilateral pleural effusions. no acute bony abnormality. no atelectasis.

Our method

there is stable cardiomegaly. no visible pneumothorax. bilateral pleural effusions identified. no suspicious pulmonary mass or nodule is seen. mild edema is seen.

Fig. 1. An example of a report generated by state-of-the-art method CMN [4] and our method. The tokens marked in blue indicate the abnormalities detected by our method but missed in CMN.

There have been several works [3], [4], [11] focusing on addressing cross-modal challenges. AlignTransformer [3] predicted the disease tags from the input image and then aligned these tags with the corresponding visual regions, thereby supplying semantic-related visual features to the decoder. CMN [4] proposed using memory networks to record cross-modal alignment and facilitate generation across modalities. However, the memory matrix is randomly initialized and lacks knowledge or useful information, which significantly impacts the clinical accuracy of report generation. Fig. 1 shows a comparison of the reports generated by the state-of-the-art method CMN [4] and our approach, showing the effective detection of interrelated diseases like cardiomegaly and edema by our proposed cross-modal knowledge-driven network.

Given the intricate and interconnected nature of abnormal regions in radiology images, disease identification becomes

*Corresponding author.

979-8-3503-3748-8/23/31.00 © 2023*IEEE*

challenging without prior medical knowledge. Thus, incorporating prior medical knowledge as complementary information is crucial for accurately reporting findings. Several studies [2], [12]–[14] have employed medical knowledge graphs to depict relationships between abnormalities, allowing the model to leverage prior knowledge and produce reports with enhanced clinical accuracy. Nevertheless, previous methods encounter challenges when integrating medical knowledge into multi-modal networks, particularly in terms of inefficient knowledge transfer across different modalities. These methods predominantly concentrate on learning single-modal features and only integrate knowledge at the encoding or decoding stage, resulting in the oversight of critical findings during report generation. Consequently, there is a lack of a method that can effectively incorporate medical knowledge and facilitate simultaneous knowledge acquisition between images and text, impeding direct and efficient cross-modal knowledge exchange while generating clinically accurate reports.

To address these concerns, in this paper, we propose an intuitive and effective strategy to enhance the clinical accuracy of the generated reports by incorporating efficient medical knowledge learning into cross-modal networks. We propose a **C**ross-modal **K**nowledge driven **Net**work (CKNet) to capture cross-modal mappings and facilitate knowledge transfer across modalities. In particular, we utilize a pre-constructed knowledge graph as a shared matrix that connects the visual and textual domains. We propose a new module that consists of two processes: knowledge distillation and preservation. These processes enable the integration of the most relevant knowledge into input images and texts, respectively. Furthermore, the shared knowledge matrix serves as a bridge connecting the two modalities, leading to enhanced efficiency in handling cross-modal interactions and facilitating smoother and more effective knowledge communication across modalities.

Our main contributions can be summarized as follows:

- We propose a new cross-modal knowledge-driven network (CKNet) and utilize a pre-constructed knowledge graph as a shared matrix to connect the two modalities.
- We propose a module comprising knowledge distillation and preservation. This module enables the integration of the most relevant knowledge into input images and texts, respectively. It facilitates smoother and more effective knowledge communication across modalities.
- Extensive experiments on two benchmarks (i.e., IU-Xray [15] and MIMIC-CXR [16]) show that our model outperforms the state-of-the-art in radiology report generation.

II. Related Work

A. Image Caption and Paragraph Generation

Image captioning involves generating a sentence that describes the input image in natural language. The prevailing architecture for the caption task is based on the encoder-decoder framework proposed by Show-Tell [9]. Building upon this framework, numerous attention mechanisms have been proposed to focus on salient visual or language signals [10], [17], [18]. Given the limited ability to describe an image in a single sentence, the task of paragraph generation has been introduced to generate a lengthy and semantically coherent paragraph based on an input image. For this purpose, the hierarchical RNN structure is commonly employed [19], [20]. Due to the limited capability of RNNs in capturing long-range dependencies, recent studies have introduced Transformer-based models [21], [22].

B. Cross-modal Radiology Report Generation

Radiology report generation, as an application and extension of image captioning to the medical domain, where most existing report generation methods follow the encoder-decoder paradigm [1], [2], [6]–[8], [12], [14], [23], [24]. Several approaches have been proposed to tackle the challenges posed by cross-modal problems and improve the clinical accuracy of generated reports. Xue et al. [23] introduced a multimodal report generation model that incorporates an iterative decoder with visual and semantic attention, aiming to enhance coherence between sentences in a recurrent manner. TieNet [25] presented an attention-encoded text embedding and saliency-weighted global average pooling approach, enabling joint learning of textual and image information and enhancing the model's capacity in describing abnormalities. To explicitly record the mappings between visual and textual modalities to facilitate report generation, R2GenCMN [4] proposed using memory networks to enhance and smooth such mapping. XPRONet [11] introduced a prototype matrix to record cross-modal prototypes and embed cross-modal information into visual and textual features, resulting in further improvements.

C. Knowledge Based Radiology Report Generation

Numerous studies have explored incorporating prior knowledge into the generation model to improve the quality of report generation. Li et al. [2] developed a report generation system incorporating knowledge-driven encoding, retrieval, and paragraphing modules. Liu et al. [14] modeled previous work experience and prior medical knowledge by emulating the working patterns of radiologists, leveraging retrieved reports and a medical knowledge graph. Zhang et al. [12] constructed a medical knowledge graph to discover abnormality relationships in report generation. PPKED [14] utilized global representations derived from pre-retrieved reports in the training corpus to model domain-specific knowledge. In contrast, our objective is to directly employ the knowledge graph as a shared matrix, bridging the gap between visual and textual modalities and addressing the challenge of inefficient knowledge transfer in cross-modal problems.

III. Methodology

A. Problem Formulation

Before introducing the cross-modal knowledge-driven network (CKNet), we present the problem formulation. Our aim is to generate a detailed radiology report $\hat{R} = \{\hat{y}_1, \hat{y}_2, \dots\}$ that describes the observations in terms of what and where based on a given radiology image I.

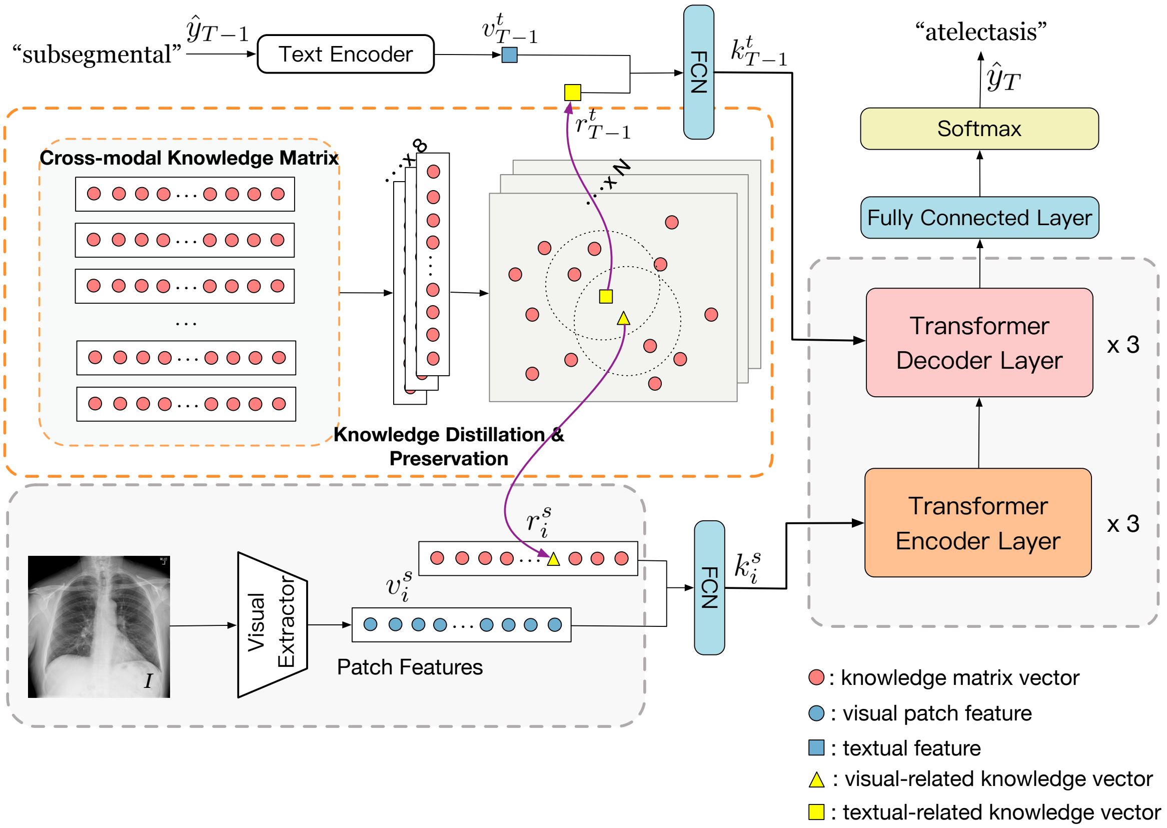


Fig. 2. The overall architecture of CKNet, where the cross-modal knowledge-driven network is illustrated in orange dash boxes. After obtaining visual patch features and textual features, they are sent to the Cross-modal Knowledge Matrix for knowledge distillation and preservation. This process incorporates relevant knowledge representations into both inputs, bringing visual and textual modalities closer together. Subsequently, the preserved knowledge representations enrich the single-modal features via a fully connected layer, serving as the source inputs for the Transformer encoder-decoder to generate the report.

The motivation for designing the cross-modal knowledge-driven network is that medical reports often necessitate specialized domain knowledge [26]. Therefore, we employ a knowledge graph that captures the structure of domain-specific knowledge to bridge the gap between visual and textual domains, facilitating effective knowledge transfer across modalities. In particular, we construct an off-the-shelf medical knowledge graph $\mathcal{G} = (V, E)$ that encompasses major clinical diseases, following the approach of [12]. Subsequently, the knowledge graph is embedded using GCN [27], taking images as input and integrating a fully-connected layer for multi-label classification. Consequently, we obtain a set of node embeddings $G = \{g_1, g_2, \ldots, g_i, \ldots, g_N\}$, where N represents the number of nodes in the graph. Each node can focus on a distinct region of the image corresponding to specific chest abnormalities.

In the subsequent modules, the parameters in the graph embeddings are fixed, and the node embeddings are utilized as a shared knowledge matrix to bridge the gap between visual and textual domains. The knowledge matrix vectors represent the spatial positions and characteristics of clinical diseases, facilitating efficient information capture across modalities and enabling effective cross-modal knowledge transmission. Importantly, the knowledge graph is trained solely on the training set of each dataset, ensuring no label leakage.

B. Basic Architecture

A visual extractor is used to initially extract visual feature maps $\boldsymbol{v} \in \mathbb{R}^{H \times W \times C}$ from radiology images. The visual feature maps $\boldsymbol{v}$ are then flattened into a sequence $\boldsymbol{v_s} \in \mathbb{R}^{(H \cdot W) \times C}$. For frontal and lateral X-rays of a patient, the visual features from these different views are concatenated to obtain the final visual representation, which serves as the input for all subsequent modules. This process is formulated as:

$$\{v_1^s, v_2^s, \ldots, v_i^s, \ldots, v_{N_s}^s\} = f_{img}(I), \tag{1}$$

where $N_s = H \times W = 49$, v_i^s denotes image patch features in the i^{th} position, and f_{img} refers to the visual extractor.

Define T as a radiology report consisting of l words, $T = \{w_1, w_2, \ldots, w_l\}$. Utilizing BioClinicalBERT [28], which is pre-trained on medical texts from the MIMIC III dataset [29], we extract text features $\boldsymbol{v_t} = \{v_1^t, v_2^t, \ldots, v_i^t, \ldots, v_l^t\}$ from the hidden state of the last layer. Here, v_i^t represents the word embedding of the i^{th} word in the report.

Then, the visual features $\boldsymbol{v_s}$ and textual features $\boldsymbol{v_t}$ are separately inputted into the cross-modal knowledge-driven network. And the obtained knowledge-aware features for visual and textual features are fed into the encoder-decoder of the

Transformer to facilitate report generation. The core of the Transformer is the Multi-Head Attention (MHA), defined as:

$$Att_m(Q,K,V) = softmax(\frac{QW_m^Q(KW_m^K)^{\mathrm{T}}}{\sqrt{p/q}})VW_m^V, \quad (2)$$

$$MHA(G,V) = [Att_1(G,V,V) \sqcup \ldots \sqcup Att_q(G,V,V)]W^O, \quad (3)$$

where Q, K, V correspond to the query, key, and value, respectively. The parameter matrices W_m^Q, W_m^K, W_m^V, W^O are associated with the m^{th} head that needs to be learned. p denotes the dimension of the input feature for each head, and q represents the number of heads in the MHA. The symbol $\sqcup$ indicates the concatenation operation, and G refers to the graph node embeddings, which are essential for bridging the two modalities. Following each of the aforementioned sub-layers, a residual connection and layer normalization are applied. Finally, the last MHA in the decoder is followed by the softmax operation.

C. Cross-modal Knowledge Driven Network

Fig. 2 illustrates the overall architecture of CKNet. We utilize node embeddings as a shared knowledge matrix that connects the visual and textual domains. This matrix effectively captures the mappings between different modalities and facilitates the transfer of knowledge across these modalities. Specifically, we propose two processes: knowledge distillation and preservation, to integrate the most relevant knowledge into visual or textual inputs. This integration aligns information from images and texts, facilitating effective knowledge transmission and significantly reducing the modality gap. These steps are executed during both the training and inference stages. During inference, all textual features are acquired through the generation process.

1) Knowledge Distillation: Given the input visual or textual features, our approach initially measures the similarity between the input single-modal representation and the shared knowledge matrix. By employing multi-head querying, we distil visual- or textual-related knowledge representations from the node embeddings.

Before feeding visual or textual features into the knowledge matrix, we perform a linear transformation to project them into the same dimension by:

$$q_i^s = v_i^s \cdot \mathbf{W}_q, \qquad q_i^t = v_i^t \cdot \mathbf{W}_q, \qquad \mathbf{k_i} = g_i \cdot \mathbf{W}_k, \quad (4)$$

where $\mathbf{W}_q$ and $\mathbf{W}_k$ are two learnable weights. Then the similarity between the input single-modal representation and the knowledge matrix is computed by:

$$D_{s_i} = \frac{q_i^s \cdot \mathbf{k_i}^\top}{\sqrt{d}}, \qquad D_{t_i} = \frac{q_i^t \cdot \mathbf{k_i}^\top}{\sqrt{d}}, \quad (5)$$

where D_{s_i} and D_{t_i} represent the visual and textual distances, respectively, between the input single-modal representation and the knowledge matrix vectors.

Afterwards, we select γ most related vectors to be preserved knowledge matrix vectors and calculate their weights w_{s_i} and w_{t_i} by normalizing the distances D_{s_i} and D_{t_i}. This process is calculated as follows:

$$w_{s_i} = \frac{D_{s_i}}{\sum_{j=1}^{\gamma} D_{s_j}}, \qquad w_{t_i} = \frac{D_{t_i}}{\sum_{j=1}^{\gamma} D_{t_j}}. \quad (6)$$

2) Knowledge Preservation: Once we have obtained the top γ similar knowledge matrix vectors and their weights, the next step is to preserve the distilled knowledge representations within the input single-modal representation. This process is also conducted in a multi-head manner. For each head, we first transform the queried knowledge matrix vectors to the same representation space as the query vectors using a fully connected layer:

$$g_s = g_i^s \cdot \mathbf{W_g}, \qquad g_t = g_i^t \cdot \mathbf{W_g}, \quad (7)$$

where g_i^s and g_i^t represent the knowledge matrix vectors which are most similar to the i^{th} image patch and word features, respectively. The transformed knowledge vectors for visual and textual features are denoted as g_s and g_t. Then we obtain the preserved knowledge vectors r_s and r_t as follows:

$$r_s = \sum_{j=1}^{\gamma} w_s \cdot g_s, \qquad r_t = \sum_{j=1}^{\gamma} w_t \cdot g_t, \quad (8)$$

where w_s and w_t are weights obtained from knowledge distillation, r_s and r_t are the preserved knowledge vectors for visual and textual features.

The last step is to fuse the preserved knowledge vectors into the input visual or textual features, respectively. This process is as follows:

$$\begin{aligned} k_s &= \mathbf{FCN}(Concat(v_s, r_s)), \\ k_t &= \mathbf{FCN}(Concat(v_t, r_t)), \end{aligned} \quad (9)$$

where $\mathbf{FCN}$ denotes the fully connected layer and $Concat$ is the concatenation operation. In conclusion, since the visual and textual features query from the same knowledge matrix, these processes narrow the gap between modalities and capture cross-modal alignment. Furthermore, the shared knowledge matrix provides prior knowledge to both the visual and textual domains. These processes facilitate effective knowledge transfer between the two modalities and encourage the decoder to generate more accurate reports. The outputs of this module serve as the source inputs for the Transformer encoder and decoder to generate the reports.

D. Report Generation via Transformer

As previously mentioned, our encoder-decoder is built based on a standard Transformer. At first, the preserved knowledge for visual features k_s is fed into the Encoder to generate intermediate states. Then the intermediate states combined with the preserved knowledge for textual features k_t are fed into Decoder to generate current output y_T. This process can be expressed as:

$$\{m_1, m_2, \ldots, m_{N_s}\} = f_e(k_{s_1}, k_{s_2}, \ldots, k_{s_{N_s}}), \quad (10)$$

$$\hat{y}_T = f_d(m_1, m_2, \ldots, m_{N_s}; k_{t_1}, k_{t_2}, \ldots, k_{t_{T-1}}), \quad (11)$$

TABLE I

THE STATISTICS OF IU X-RAY AND MIMIC-CXR. IMAGES, REPORTS, PATIENTS AND AVG. LEN. REPRESENT THE NUMBER OF IMAGES, REPORTS, PATIENTS, AND THE AVERAGE LENGTH OF REPORTS.

Dataset	Split	#Images	#Reports	#Patients	Avg. Len.
IU-Xray [15]	Train	5,212	2,780	2,780	38.29
	Val	720	402	402	36.58
	Test	1,534	800	800	37.63
MIMIC-CXR [29]	Train	368,960	222,758	64,586	53.00
	Val	2,991	1,808	500	53.05
	Test	5,159	3,269	293	66.40

where f_e and f_d refers to the encoder and decoder, y_T denotes the word prediction for time step T. The above process is repeated until the complete report is generated.

IV. EXPERIMENTS

A. Datasets

The experiments are performed on two publicly available datasets IU-Xray [15] and MIMIC-CXR [16]. Following the same data splits as [5], [11], [30], we divide the IU-Xray dataset into train (70%), validation (10%) and test (20%) sets and remove samples without both views of images. For MIMIC-CXR, we adopt its official split [16]. There is no overlap of patients across the train, validation and test sets. Table I shows the statistics of the two datasets.

B. Evaluation Metrics

We evaluate the performance of models using both conventional natural language generation (NLG) metrics and clinical efficacy (CE) metrics. The NLG metrics used for evaluation include BLEU [32], METEOR [33], ROUGE-L [34] and CIDEr [35]. BLEU-1 to BLEU-4 scores are calculated based on consecutive words in the prediction report. ROUGE-L measures recall of consecutive word sequences, while BLEU-n calculates accuracy. The CIDEr score evaluates the coverage of essential information in the generated text compared to the ground truth. The METEOR indicator primarily considers word overlap and lexical abbreviation expansion, allowing for the incorporation of additional syntactic and semantic information. To evaluate how well the generated reports describe abnormalities, we further report CE metrics following previous work [5]. For the CE metrics, we use the CheXpert [36] labeler to extract labels from the generated reports and compare the results with ground truth for 14 different thoracic diseases using accuracy, precision, recall, F1 and AUC. Since the IU-Xray dataset lacks consistent labels, we only report CE metrics for the MIMIC-CXR dataset.

C. Implementation Details

For each dataset, we first obtain the corresponding trained knowledge graph embeddings following the approach in [12]. The knowledge graph consists of 40 nodes and a dimension of 512. The graph node embeddings serve as a shared knowledge matrix, connecting two domains and facilitating the transmission of knowledge. Importantly, the graph weights are kept frozen throughout the training and inference stages of report generation.

To ensure consistency with the experiment settings of previous work [5], [11], we utilize both the frontal and lateral radiology images of a patient on IU-Xray by concatenating the visual features, and one image for MIMIC-CXR. We adopt DenseNet-121 [37] pre-trained on CheXpert [36] as our visual extractor. The extracted features are $2,048$ feature maps in the shape of 7×7 which are further projected into 512 feature maps, i.e., N_s is 49 and C is 512. For the text encoder, we use BioClinicalBERT [28] pre-trained on medical texts from the MIMIC-III dataset [29] to extract text features, and the global textual representation has 768 dimensions.

For the encoder-decoder backbone, we use a Transformer structure with 3 layers and 8 attention heads, and the dimension of hidden states is 512. According to the report generation performance on the validation set, the number of most related vectors γ is set to 8, meaning that only the top 8 knowledge matrix vectors are selected to merge with the single-modal representations. We train our model under cross-entropy loss with Adam optimizer [38]. The learning rates are set to 1×10^{-3} and 5×10^{-4} for the visual extractor and encoder-decoder on IU-Xray, while MIMIC-CXR has a smaller learning rate with 5×10^{-5} and 1×10^{-4} respectively. We decay them by 0.8 rate per epoch and the bath sizes are 16 for all datasets. To balance effectiveness and efficiency, we adopt a beam size of 3 in the report generation process. Note that the optimal hyper-parameters are determined by estimating the models on the validation sets and we report the results on the testing set when the validation set achieves the best BLEU-4 score. Our model is implemented using the PyTorch [39] deep learning framework.

D. Quantitative Results

1) Language Generation Performance: We compare our proposed model with previous state-of-the-art methods, i.e., CoATT [1], KERP [2], SentKG [12], R2Gen [5], R2GenCMN [4], PPKED [14] and CMCL [31]. The NLG results are shown in Table II. It is evident that our proposed CKNet surpasses the baselines in nearly all metrics, suggesting that our method emphasizes overall contextual information and excels in capturing longer n-grams. Additionally, our model achieves significantly higher CIDEr and ROUGE-L scores compared to the baselines, signifying its ability to generate more cognitively fluent sentences. Remarkably, our method outperforms R2GenCMN by 14.5%, 15.7% and 14.1% on the BLEU-2, BLEU-3 and BLEU-4 scores on IU-Sray dataset, respectively. A similar improvement can be observed on the MIMIC-CXR benchmark. Meanwhile, we notice that our model's improvement on MIMIC-CXR is not as substantial as on IU-Xray, possibly due to the larger dataset posing challenges for knowledge learning. Nonetheless, CKNet can still achieve improvements on BLEU-2 to BLEU-4, indicating

TABLE II
THE PERFORMANCES OF OUR MODEL COMPARED WITH PREVIOUS STUDIES ON THE TEST SETS OF IU-XRAY AND MIMIC-CXR. THE BOLD SCORES INDICATE THE BEST RESULTS FOR EACH METRIC.

Dataset	Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr
IU-Xray [15]	CoATT [1]	0.455	0.288	0.205	0.154	-	0.369	0.277
	SentKG [12]	0.441	0.291	0.203	0.147	-	0.367	0.304
	KERP [2]	0.482	0.325	0.226	0.162	-	0.339	0.280
	R2Gen [5]	0.470	0.304	0.219	0.165	0.187	0.371	-
	PPKED [14]	0.483	0.315	0.224	0.168	-	0.376	0.351
	R2GenCMN [4]	0.475	0.309	0.222	0.170	0.191	0.375	-
	CKNet (ours)	**0.515**	**0.354**	**0.257**	**0.194**	**0.213**	**0.402**	**0.392**
MIMIC-CXR [16]	CoATT [1]	0.331	0.220	0.147	0.117	-	0.276	-
	R2Gen [5]	0.353	0.218	0.145	0.103	0.142	0.277	-
	CMCL [31]	0.334	0.217	0.140	0.097	-	0.281	
	PPKED [14]	**0.360**	0.224	0.149	0.106	**0.149**	0.284	-
	R2GenCMN [4]	0.353	0.218	0.148	0.106	0.142	0.278	-
	CKNet (ours)	0.356	**0.239**	**0.157**	**0.113**	0.146	**0.289**	**0.118**

TABLE III
THE RESULTS OF CLINICAL EFFICACY METRICS ON THE TEST SET OF THE MIMIC-CXR DATASET. FOR THE BASELINES MARKED BY *, WE REPLICATE THE EXPERIMENTS BY RUNNING THEIR CODES.

Model	Accuracy	Precision	Recall	F1 score	AUC
ST* [9]	0.204	0.244	0.197	0.190	0.685
AdaAtt* [40]	0.217	0.266	0.192	0.195	0.701
Att2In* [41]	0.232	0.310	0.225	0.234	0.732
TopDown* [10]	0.228	0.312	0.235	0.244	0.734
R2Gen [5]	0.297	0.333	0.273	0.276	0.763
R2GenCMN [4]	0.321	0.334	0.275	0.278	0.775
CKNet (Ours)	**0.368**	**0.423**	**0.348**	**0.358**	**0.802**

that the knowledge matrix used to bridge the gap between modalities can improve knowledge learning efficiency.

2) Clinical Accuracy Performance: To further demonstrate the effectiveness of our model on clinical efficacy (CE) metrics, we compare it with conventional image captioning works, e.g., ST [9], AdaAtt [40], Att2In [41], and TopDown [10], as well as specific medical report generation methods R2Gen [5] and R2GenCMN [4]. Since the IU-Xray dataset does not provide consistent labels, we solely report CE metrics on the MIMIC-CXR dataset. The CE metrics can help to evaluate the accuracy of the generated reports in describing abnormalities. The results in Table III reveal that our proposed method exhibits a significant superiority over all prior models in terms of CE metrics, leading to a 27.2% rise in precision, 26.5% in recall, and 28.8% in the F1 score. This implies that from the clinical perspective, our model has produced more accurate reports than other models. Two potential reasons can be identified. Firstly, the adoption of a cross-modal knowledge matrix bridges the gap between the two modalities, enabling more intuitive and efficient knowledge transfer between texts and images. Secondly, the proposed knowledge distillation and preservation effectively capture medical knowledge and cross-modal mappings, assisting in report generation.

E. Qualitative Results

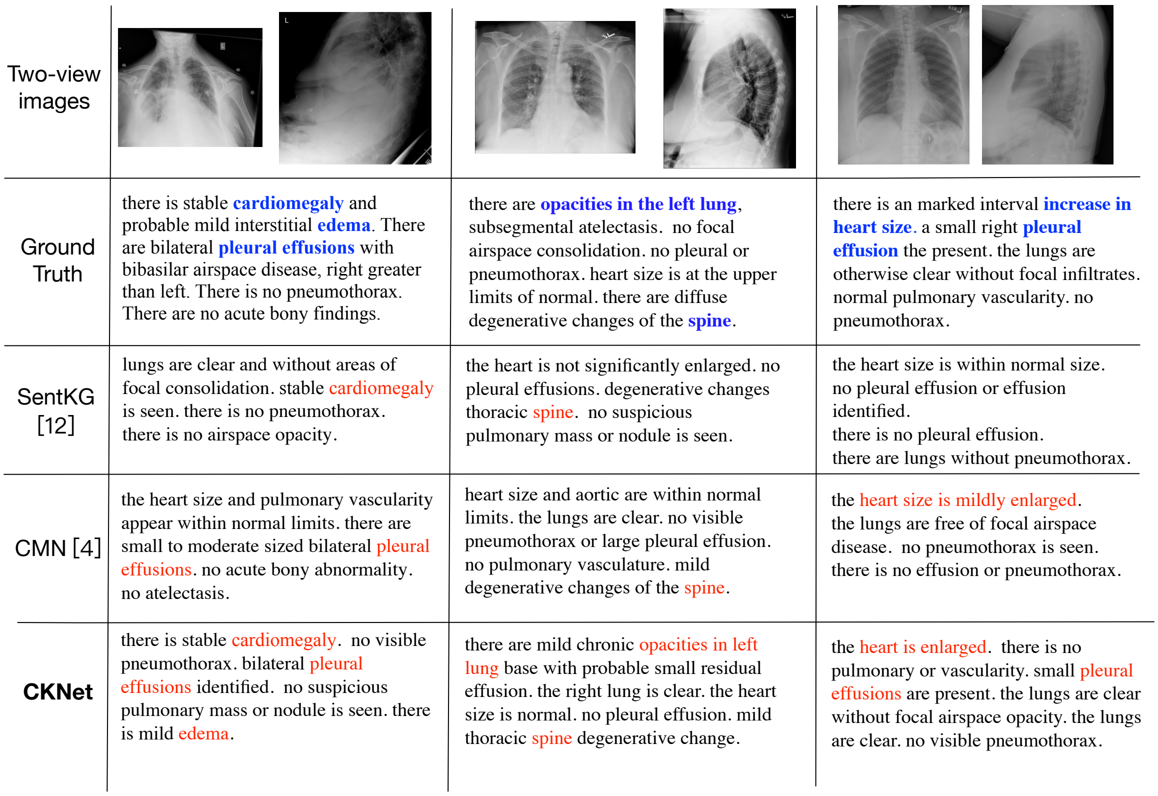
We further conduct a qualitative analysis to gain a perceptual understanding of the improvements. Fig. 3 presents three cases of ground-truth reports and reports generated by our method CKNet and the state-of-the-art baselines. It is evident that CKNet accurately generates descriptions of abnormalities or diseases in all three cases. Notably, in the second column, CKNet successfully identifies both opacity in the left lung and spine, whereas other methods fail to detect either of them, confirming its superiority in generating reports of higher quality. This suggests that the cross-modal knowledge interactions proposed, based on the shared knowledge matrix, offer more effective knowledge for report generation and enhance the clinical accuracy of the generated reports.

F. Ablation Study

We conduct experiments to verify the effectiveness of the proposed method components, as displayed in Table IV. For the first setting (Base), we remove other modules and only use the visual extractor (DenseNet-121) and encoder-decoder (Transformer) backbone. For the second setting (+MEN), we replace CKNet with a matrix of the same dimension and initialize it randomly without knowledge. For the third setting (+CKNet), we implement the full CKNet with all proposed components. A noticeable decrease occurs in every metric when the shared knowledge matrix is not employed, showcasing its role in facilitating the learning of abnormality-related knowledge and bridging the two modalities through the alignment of cross-modal representations. Furthermore, we observe a higher performance increase on IU-Xray compared

TABLE IV
THE EXPERIMENTAL RESULTS OF ABLATION STUDIES ON THE IU-XRAY AND MIMIC-CXR DATASETS. THE BEST VALUES ARE HIGHLIGHTED IN BOLD.

Dataset	Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr
IU-Xray [15]	Base	0.445	0.293	0.205	0.144	0.178	0.355	0.343
	+MEN	0.472	0.323	0.229	0.168	0.192	0.376	0.366
	+CKNet	**0.515**	**0.354**	**0.257**	**0.194**	**0.213**	**0.402**	**0.392**
MIMIC-CXR [16]	Base	0.307	0.199	0.119	0.088	0.113	0.253	0.089
	+MEN	0.329	0.212	0.125	0.092	0.125	0.269	0.102
	+CKNet	**0.356**	**0.239**	**0.157**	**0.113**	**0.146**	**0.289**	**0.118**



Two-view images			
Ground Truth	there is stable **cardiomegaly** and probable mild interstitial **edema**. There are bilateral **pleural effusions** with bibasilar airspace disease, right greater than left. There is no pneumothorax. There are no acute bony findings.	there are **opacities in the left lung**, subsegmental atelectasis. no focal airspace consolidation. no pleural or pneumothorax. heart size is at the upper limits of normal. there are diffuse degenerative changes of the **spine**.	there is an marked interval **increase in heart size**. a small right **pleural effusion** the present. the lungs are otherwise clear without focal infiltrates. normal pulmonary vascularity. no pneumothorax.
SentKG [12]	lungs are clear and without areas of focal consolidation. stable cardiomegaly is seen. there is no pneumothorax. there is no airspace opacity.	the heart is not significantly enlarged. no pleural effusions. degenerative changes thoracic spine. no suspicious pulmonary mass or nodule is seen.	the heart size is within normal size. no pleural effusion or effusion identified. there is no pleural effusion. there are lungs without pneumothorax.
CMN [4]	the heart size and pulmonary vascularity appear within normal limits. there are small to moderate sized bilateral pleural effusions. no acute bony abnormality. no atelectasis.	heart size and aortic are within normal limits. the lungs are clear. no visible pneumothorax or large pleural effusion. no pulmonary vasculature. mild degenerative changes of the spine.	the heart size is mildly enlarged. the lungs are free of focal airspace disease. no pneumothorax is seen. there is no effusion or pneumothorax.
CKNet	there is stable cardiomegaly. no visible pneumothorax. bilateral pleural effusions identified. no suspicious pulmonary mass or nodule is seen. there is mild edema.	there are mild chronic opacities in left lung base with probable small residual effusion. the right lung is clear. the heart size is normal. no pleural effusion. mild thoracic spine degenerative change.	the heart is enlarged. there is no pulmonary or vascularity. small pleural effusions are present. the lungs are clear without focal airspace opacity. the lungs are clear. no visible pneumothorax.

Fig. 3. Three cases of ground-truth reports and reports generated by our method CKNet and the state-of-the-art baselines. Tokens marked in blue indicate the presence of abnormalities or diseases in the ground-truth reports, while those marked in red indicate the accurate detection of abnormalities or diseases in the generated reports.

to MIMIC-CXR, signifying that different from scenes with large labeled sample sizes, scenarios with small data amounts require more assistance from knowledge.

V. CONCLUSION

In this paper, we presented CKNet, a new cross-modal knowledge-driven network for radiology report generation. We utilized a pre-constructed knowledge graph as a shared matrix to bridge the gap between visual and textual modalities, facilitating cross-modal knowledge transfer. On the one hand, the shared knowledge matrix captures cross-modal mappings, reducing the disparity between images and texts. On the other hand, the proposed new module for knowledge distillation and preservation integrates relevant knowledge representations into both visual and textual inputs, enhancing the intuitive cross-modal knowledge interaction and improving the clinical accuracy of generated reports. Experimental results on two benchmark datasets demonstrated the effectiveness of our proposed CKNet, which outperforms state-of-the-art methods.

ACKNOWLEDGMENT

This work is partially supported by the National Key Research and Development Plan Project 2022YFC3600901, the Chinesisch-Deutsche Kooperationsgruppe: Precision Medicine in Pancreatic Cancer Antrag GZ 1456, and the Royal Society grant IES\R3\223050.

References

[1] B. Jing, P. Xie, and E. Xing, "On the automatic generation of medical imaging reports," *arXiv preprint arXiv:1711.08195*, 2017.

[2] C. Y. Li, X. Liang, Z. Hu, and E. P. Xing, "Knowledge-driven encode, retrieve, paraphrase for medical image report generation," in *AAAI*, vol. 33, no. 01, 2019, pp. 6666–6673.

[3] D. You, F. Liu, S. Ge, X. Xie, J. Zhang, and X. Wu, "Aligntransformer: Hierarchical alignment of visual regions and disease tags for medical report generation," in *MICCAI*, 2021, pp. 72–82.

[4] Z. Chen, Y. Shen, Y. Song, and X. Wan, "Cross-modal memory networks for radiology report generation," in *ACL*, 2021, pp. 5904–5914.

[5] Z. Chen, Y. Song, T.-H. Chang, and X. Wan, "Generating radiology reports via memory-driven transformer," *arXiv preprint arXiv:2010.16056*, 2020.

[6] X. Xie, Y. Xiong, P. S. Yu, K. Li, S. Zhang, and Y. Zhu, "Attention-based abnormal-aware fusion network for radiology report generation," in *Database Systems for Advanced Applications: DASFAA 2019 International Workshops: BDMS, BDQM, and GDMA, Chiang Mai, Thailand, April 22–25, 2019, Proceedings 24*. Springer, 2019, pp. 448–452.

[7] X. Jia, Y. Xiong, J. Zhang, Y. Zhang, and Y. Zhu, "Few-shot radiology report generation for rare diseases," in *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2020, pp. 601–608.

[8] X. Jia, Y. Xiong, J. Zhang, Y. Zhang, B. Suzanne, Y. Zhu, and C. Tang, "Radiology report generation for rare diseases via few-shot transformer," in *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2021, pp. 1347–1352.

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

[10] 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 *CVPR*, 2018, pp. 6077–6086.

[11] J. Wang, A. Bhalerao, and Y. He, "Cross-modal prototype driven network for radiology report generation," in *ECCV*. Springer, 2022, pp. 563–579.

[12] Y. Zhang, X. Wang, Z. Xu, Q. Yu, A. Yuille, and D. Xu, "When radiology report generation meets knowledge graph," in *AAAI*, vol. 34, no. 07, 2020, pp. 12 910–12 917.

[13] Y. Cao, L. Cui, F. Yu, L. Zhang, Z. Li, N. Liu, and Y. Xu, "Kdtnet: Medical image report generation via knowledge-driven transformer," in *DASFAA*, 2022, pp. 117–132.

[14] F. Liu, X. Wu, S. Ge, W. Fan, and Y. Zou, "Exploring and distilling posterior and prior knowledge for radiology report generation," in *CVPR*, 2021, pp. 13 753–13 762.

[15] D. Demner-Fushman, M. D. Kohli, M. B. Rosenman, S. E. Shooshan, L. Rodriguez, S. Antani, G. R. Thoma, and C. J. McDonald, "Preparing a collection of radiology examinations for distribution and retrieval," *Journal of the American Medical Informatics Association*, vol. 23, no. 2, pp. 304–310, 2016.

[16] A. E. Johnson, T. J. Pollard, N. R. Greenbaum, M. P. Lungren, C.-y. Deng, Y. Peng, Z. Lu, R. G. Mark, S. J. Berkowitz, and S. Horng, "Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs," *arXiv preprint arXiv:1901.07042*, 2019.

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

[18] Z.-J. Zha, D. Liu, H. Zhang, Y. Zhang, and F. Wu, "Context-aware visual policy network for fine-grained image captioning," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 2, pp. 710–722, 2019.

[19] J. Krause, J. Johnson, R. Krishna, and L. Fei-Fei, "A hierarchical approach for generating descriptive image paragraphs," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 317–325.

[20] J. Wang, Y. Pan, T. Yao, J. Tang, and T. Mei, "Convolutional auto-encoding of sentence topics for image paragraph generation," *arXiv preprint arXiv:1908.00249*, 2019.

[21] M. Cornia, M. Stefanini, L. Baraldi, and R. Cucchiara, "Meshed-memory transformer for image captioning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 578–10 587.

[22] G. Li, L. Zhu, P. Liu, and Y. Yang, "Entangled transformer for image captioning," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 8928–8937.

[23] Y. Xue, T. Xu, L. Rodney Long, Z. Xue, S. Antani, G. R. Thoma, and X. Huang, "Multimodal recurrent model with attention for automated radiology report generation," in *MICCAI*, 2018, pp. 457–466.

[24] X. Jia, Y. Xiong, J. Zhang, Y. Zhang, Y. Zhu, and S. Y. Philip, "Few-shot radiology report generation via knowledge transfer and multi-modal alignment," in *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2022, pp. 1574–1579.

[25] X. Wang, Y. Peng, L. Lu, Z. Lu, and R. M. Summers, "Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays," in *CVPR*, 2018, pp. 9049–9058.

[26] S. K. Goergen, F. J. Pool, T. J. Turner, J. E. Grimm, M. N. Appleyard, C. Crock, M. C. Fahey, M. F. Fay, N. J. Ferris, S. M. Liew *et al.*, "Evidence-based guideline for the written radiology report: Methods, recommendations and implementation challenges," *Journal of medical imaging and radiation oncology*, vol. 57, no. 1, pp. 1–7, 2013.

[27] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.

[28] E. Alsentzer, J. R. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann, and M. McDermott, "Publicly available clinical bert embeddings," *arXiv preprint arXiv:1904.03323*, 2019.

[29] A. E. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, and R. G. Mark, "Mimic-iii, a freely accessible critical care database," *Scientific data*, vol. 3, no. 1, pp. 1–9, 2016.

[30] Y. Li, X. Liang, Z. Hu, and E. P. Xing, "Hybrid retrieval-generation reinforced agent for medical image report generation," *Advances in neural information processing systems*, vol. 31, 2018.

[31] F. Liu, S. Ge, and X. Wu, "Competence-based multimodal curriculum learning for medical report generation," in *ACL/IJCNLP (1)*, 2021.

[32] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *ACL*, 2002, pp. 311–318.

[33] M. Denkowski and A. Lavie, "Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems," in *WMT*, 2011, pp. 85–91.

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

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

[36] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya *et al.*, "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 590–597.

[37] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[38] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[39] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.

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

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