A Systematic Review of Deep Learning-based Research on Radiology Report Generation

Chang Liu[♠], Yuanhe Tian[♠♥], and Yan Song[♠*]

[♠]University of Science and Technology of China [♥]University of Washington

[♠]lc980413@mail.ustc.edu.com [♥]yhtian@uw.edu [♠]clksong@gmail.com

Abstract—Radiology report generation (RRG) aims to automatically generate free-text descriptions from clinical radiographs, e.g., chest X-Ray images. RRG plays an essential role in promoting clinical automation and presents significant help to provide practical assistance for inexperienced doctors and alleviate radiologists' workloads. Therefore, consider these meaningful potentials, research on RRG is experiencing explosive growth in the past half-decade, especially with the rapid development of deep learning approaches. Existing studies perform RRG from the perspective of enhancing different modalities, provide insights on optimizing the report generation process with elaborated features from both visual and textual information, and further facilitate RRG with the cross-modal interactions among them. In this paper, we present a comprehensive review of deep learning-based RRG from various perspectives. Specifically, we firstly cover pivotal RRG approaches based on the task-specific features of radiographs, reports, and the cross-modal relations between them, and then illustrate the benchmark datasets conventionally used for this task with evaluation metrics, subsequently analyze the performance of different approaches and finally offer our summary on the challenges and the trends in future directions. Overall, the goal of this paper is to serve as a tool for understanding existing literature and inspiring potential valuable research in the field of RRG.†

Index Terms—Radiology Report Generation, Deep Learning, Survey.

1 Introduction

With the prosperity of artificial intelligence (AI), medical industry has emerged a growing demand for its advanced techniques, where different medical applications in urgent needs of automatically processing massive medical data. These applications usually comprise multiple research directions, including pathological image analysis, medical report generation, disease prediction and diagnosis, etc. In performing the aforementioned research, AI techniques are expected to handle medical data in various modalities. Particularly, medical imaging (e.g., X-Ray, MRI) serves as the mainstream medical data that are collected by different medical instruments, which presents the patients' health condition straightforwardly through the various image types. Besides image, text is another important medium playing a pivotal carrier in doctors' daily routines and medical researches, which stores various medical contents through the formulation of natural language, e.g., medical records, research papers, reports. Both image and text synergistically perform in clinical diagnosis and treatment recommendation, where, particularly, medical imaging usually requires physicians to write reports according to the syndromes of patients that are reflected in the image so as to form professional records for later processing. Among all applications requiring such connection between image and text, automatically performing radiology report generation (RRG) serves as the most practical and extensive support in the imaging department of many hospitals for lightening radiologists' burdens and assisting inexperienced physicians. Technically, RRG aims to generate descriptive texts for a radiograph, which requires the produced texts to accurately depict overall and local health condition (e.g., regional lesions, disease of specific organs), meanwhile maintaining the consistency of the entire report to the input image thus formulating the final diagnosis. Thus this task holds great significance of applying AI techniques to the medical industry, so that it emerges as an attractive research direction in AI and clinical medicine in recent years.

Findings:
Heart size is within normal limits. There are surgical clips in the left mediastinum. There is no pneumothorax. There is a small left pleural effusion. Abnormal convexity within the mediastinum represents adenopathy.

Impression:
Small left pleural effusion. Mediastinal adenopathy noted.

Fig. 1. A representative chest radiology image with its corresponding doctor-written radiology report. Specifically, each report consists of a "Findings" and a "Impression" section, where the "Findings" section records detailed descriptions by radiologists, and the "Impression" section presents an overall diagnostic summarization based on the radiograph. RRG is normally aiming at generating the "Findings" content.

To better understand RRG, Figure 1 presents an example of a chest radiology image with its description including doctor-written findings and impression. The findings section consists of detailed descriptions recorded by radiologists for each specific region of the radiology image, while the impression section contains the most information summarized from the findings section for the final clinical diagnosis. Normally RRG is targeted to generate the findings section, i.e., radiology report, according to an input radiograph. To perform RRG, early approaches produce

* Corresponding author.

† We maintain a GitHub project at https://github.com/synlp/RRG-Review to summarize RRG-related papers and resources, which is under continuously updating to facilitate future research in this field.

medical labels (e.g., locations of lesions and severities) to assist the generation process, where such approaches consist of two main categories, i.e., retrieval-based [1], [2] and disease classification-based approaches [3], [4], respectively. Particularly, the former category aims to correlate the visual features of radiographs to specific texts (i.e, disease attributes) in the report, and obtain relevant texts through retrieving a medical database. The latter category manages to directly predict medical labels from the input radiograph. For both categories of the approaches, however, they still rely on professional radiologists to complete the writing of the entire report thus fail to achieve full automation in practical clinic environment, which drives the expectation of more effective approaches to handle this task.

Recently, with the rising of deep learning and neural networks, significant breakthrough has been witnessed in RRG research, where studies have dramatically grown in the past five years, especially with the trend presenting an explosive curve in the recent two years. Upon training on large-scale radiograph-report pairs in prevailing radiology datasets, neural approaches [5], [6], [7], [8], [9], [10], [11], [12], [13] are capable of generating the entire contents of radiology reports in an end-to-end manner. These approaches are mostly based on the encoder-decoder architecture, which normally employ visual encoders to extract visual representations from radiographs, and adopt text decoders to generate descriptive texts of radiology reports according to the extracted representations. Moreover, they also further consider task-specific features of RRG from the aspects of different modalities [10], [14], [15], [16], [17], [18], [19], [20], [21], [22], [23], [24]. Specifically, some of them emphasize on properties of radiographs or reports separately [14], [15], [16], [17], [20], [22], [23], [24], [25], [26], while others focus on establishing the vision-language connections between radiographs and reports [18], [19], [21], [27], [28], [29], [30]. All these approaches have achieved great success on the RRG task, and provide interesting findings, guidance as well as references to push this research area forward.

In this paper, we present an up-to-date and comprehensive overview of deep learning-based RRG, with particular emphases on different algorithms, datasets, and evaluation standards. We start from introducing background knowledge of the RRG task, and demonstrate our classification criteria to categorize existing deep learning-based RRG approaches based on the perspectives of enhancing corresponding modalities, including visual-only, textual-only, as well as cross-modal approaches. Then we present prevailing RRG datasets in multiple domains, including benchmarks for standard evaluation for RRG, and domain-specific datasets that emphasize on radiology studies of particular organs (e.g., CT imaging of COVID patients, liver lesions), etc. Afterwards, we summarize the evaluation principles and a series of metrics that measure the performance of RRG from different perspectives. Subsequently we analyze experimental results from different approaches based on the aforementioned evaluation metrics, and analyze the impacts of different network architectures on RRG performance. Furthermore, we discuss the challenges and trends in future directions to extend current RRG research, from perspectives of model design, modality enhancement, data augmentation, as well as evaluation metrics.

2 Approaches

RRG in general follows a cross-modal generation paradigm similar to image captioning tasks [14], [31], [32], [33], [34], [35], yet different in the resulting text length. Formally, given a radiology image $\mathcal{I}$, RRG generates its corresponding radiology report $\widehat{\mathcal{R}}$, with distilling key semantic information from $\mathcal{I}$, and accurately produce descriptive texts in $\widehat{\mathcal{R}}$. Existing approaches generally adopt the encoder-decoder architecture as their foundation pipeline, which employs a visual encoder f_v to extract high-level semantics (e.g., latent representations, medical terms, or semantic graphs) that are represented by $\mathcal{H}^v$ from $\mathcal{I}$, and utilizes a text decoder f_t to transfer $\mathcal{H}^v$ into descriptive texts in $\widehat{\mathcal{R}}$, with the overall process formulated by

$$\widehat{\mathcal{R}} = f_t\left(\mathcal{H}^v\right), \;\; \mathcal{H}^v = f_v\left(\mathcal{I}\right) \tag{1}$$

It is worth noting that most existing approaches generate $\widehat{\mathcal{R}}$ based on a single radiograph $\mathcal{I}$; although there are some approaches that process multiple radiographs with support from particular datasets, they use the concatenation of these radiographs to form them into a single input image and thus fit into our task formulation. In learning the aforementioned pipeline, existing studies utilize radiology reports $\mathcal{R}^*$ written by professional radiologists as the reference for their systems, whose output $\widehat{\mathcal{R}}$ is expected to be as close to $\mathcal{R}^*$ as possible. Particularly, they follows standard optimization procedure by generative neural models by utilizing the cross-entropy loss $\mathcal{L}_{CE}$ through comparing system output $\widehat{\mathcal{R}}$ with the gold standard report $\mathcal{R}^*$. In detail, let $\widehat{\mathcal{R}} = \{\widehat{r}_1 \dots \widehat{r}_{N_r}\}$ and $\mathcal{R}^* = \{r^*_1 \dots r^*_{N_{r^*}}\}$ with N_r and N_{r^*} denoting text token numbers in $\widehat{\mathcal{R}}$ and $\mathcal{R}^*$, respectively. To generate the i-th token, the RRG model takes the radiograph $\mathcal{I}$ and historical generated tokens (which are $r^*_1 \dots r^*_{i-1}$ in training and $\widehat{r}_1 \dots \widehat{r}_{i-1}$ in inference) and computes the probability distribution over the vocabulary $\mathcal{V}$ where the probability of generating the j-th token v_j in the vocabulary is denoted as $p_{i,j}$. Then, RRG model selects the token with the highest probability as the generated token $\widehat{r}_i = arg\,max_{v_j \in \mathcal{V}}(p_{i,j})$. Afterwards, $\widehat{r}_i$ is compared with r^*_i to calculate the cross-entropy loss $\mathcal{L}^i_{CE}$ of the i-th token by

$$\mathcal{L}^i_{CE} = -\sum_{v_j \in \mathcal{V}} p^*_{i,j} \log p_{i,j} \tag{2}$$

where $p^*_{i,j}$ is a probability defined by $p^*_{i,j} = 1$ if $v_j = r^*_i$ and $p^*_{i,j} = 0$ otherwise. Thus, $\mathcal{L}_{CE}$ over all tokens of $\widehat{\mathcal{R}}$ is obtained through $\mathcal{L}_{CE} = \sum_{i=1}^{N_r} \mathcal{L}^i_{CE}$.

Owing to the cross-modal characteristics of RRG, existing approaches put emphasis on facilitating RRG with its task-specific features from different modalities, where they consider the features from radiographs, reports, as well as the cross-modal correlations between them. Therefore, in this paper, we categorize existing approaches into three main streams based on the modality that they focus on, including visual-only, textual-only, and cross-modal approaches, respectively, where Table 1 presents such categorization on different approaches.

2.1 Visual-only Approaches

Serving as the input of RRG, radiographs are generally the only information input to the task, therefore becomes

TABLE 1
Categorization of existing RRG approaches based on our classification criteria of enhancing different modalities, including visual-only approaches, textual-only approaches, and cross-modal approaches, which are highlighted in red, green, and blue background, respectively. Additionally, we list the model architecture and evaluation standard of the surveyed studies. "GCN" denotes graph convolutional networks.

		Model Architecture		Evaluation Standard	
Study	**Method Category**	Encoder	Decoder	Datasets	Evaluation Metrics
Wang *et al.* [36]	visual-only	Transformer	Transformer	IU X-Ray, MIMIC-CXR	NLG metrics
Tanida *et al.* [23]	visual-only	CNN	Transformer	MIMIC-CXR	NLG metrics, CE metrics
Li *et al.* [24]	visual-only	GCN	Transformer	CX-CHR, COV-CTR	NLG metrics, CIDEr
Xue *et al.* [37]	textual-only	CNN	LSTM	IU X-Ray	NLG metrics
Yuan *et al.* [38]	textual-only	CNN	LSTM	IU X-Ray	NLG metrics
Jing *et al.* [14]	textual-only	LSTM	LSTM	IU X-Ray, CX-CHR	NLG metrics, CIDEr
Li *et al.* [15]	textual-only	Transformer	Transformer	IU X-Ray, CX-CHR	NLG metrics, CIDEr
Harzig *et al.* [39]	textual-only	CNN	LSTM	IU X-Ray	NLG metrics
Zhang *et al.* [16]	textual-only	GCN	LSTM	IU X-Ray, CX-CHR	NLG metrics, CIDEr, MIRQI
Chen *et al.* [8]	textual-only	Transformer	Transformer	IU X-Ray, MIMIC-CXR	NLG metrics, CE Metrics
Liu *et al.* [17]	textual-only	Transformer	Transformer	IU X-Ray, MIMIC-CXR	NLG metrics
Yan *et al.* [27]	textual-only	Transformer	Transformer	MIMIC-ABN, MIMIC-CXR	NLG metrics, CE Metrics
Liu *et al.* [40]	textual-only	Transformer	Transformer	IU X-Ray, MIMIC-CXR	NLG metrics
Nooralahzadeh *et al.* [9]	textual-only	Transformer	Transformer	IU X-Ray, MIMIC-CXR	NLG metrics, CE Metrics
You *et al.* [10]	textual-only	Transformer	Transformer	IU X-Ray	NLG metrics
Nishino *et al.* [20]	textual-only	CNN	Transformer	JLiverCT, MIMIC-CXR	NLG metrics, CE metrics, CO
Dalla Serra *et al.* [41]	textual-only	Transformer	Transformer	MIMIC-CXR	NLG metrics, CE metrics
Kale *et al.* [25]	textual-only	GCN	Transformer	IU X-Ray, MIMIC-CXR	NLG metrics, BERTScore
Yan *et al.* [42]	textual-only	CNN	Transformer	IU X-Ray, MIMIC-CXR	NLG metrics, CE Metrics
Li *et al.* [43]	textual-only	Transformer	Transformer	IU X-Ray, MIMIC-CXR	NLG metrics, CE metrics, CIDEr
Hou *et al.* [26]	textual-only	Transformer	Transformer	IU X-Ray, MIMIC-CXR	NLG metrics, CE metrics
Kale *et al.* [22]	textual-only	CNN	Transformer	IU X-Ray, MIMIC-CXR	NLG metrics, CE metrics, CIDEr
Zhang *et al.* [44]	textual-only	CNN	Transformer	IU X-Ray, MIMIC-CXR	NLG metrics, CE metrics
Jing *et al.* [5]	cross-modal	CNN	LSTM	IU X-Ray, PEIR Gross	NLG metrics
Nishino *et al.* [7]	cross-modal	RNN	GRU	CX-CHR	NLG metrics
Wang *et al.* [28]	cross-modal	CNN	LSTM	IU X-Ray, COV-CTR	NLG metrics, CIDEr
Liu *et al.* [29]	cross-modal	CNN	LSTM	IU X-Ray, MIMIC-CXR	NLG metrics, CE metrics
Ni *et al.* [19]	cross-modal	CNN	LSTM	IU X-Ray, MIMIC-CXR	NLG metrics, CIDEr, nKTD
Chen *et al.* [18]	cross-modal	Transformer	Transformer	IU X-Ray, MIMIC-CXR	NLG metrics, CE Metrics
Hou *et al.* [45]	cross-modal	CNN	Transformer	IU X-Ray, MIMIC-CXR	NLG metrics, CE metrics, CIDEr
You *et al.* [46]	cross-modal	Transformer	Transformer	IU X-Ray, MIMIC-CXR	NLG metrics
Qin *et al.* [21]	cross-modal	Transformer	Transformer	IU X-Ray, MIMIC-CXR	NLG metrics, CE Metrics
Wang *et al.* [47]	cross-modal	Transformer	Transformer	IU X-Ray, MIMIC-CXR	NLG metrics
Delbrouck *et al.* [11]	cross-modal	CNN	Transformer	IU X-Ray, MIMIC-CXR	NLG metrics, CE metrics, CIDEr
Yu *et al.* [48]	cross-modal	Transformer	Transformer	IU X-Ray, MIMIC-CXR	NLG metrics, CE metrics, CIDEr
Yan *et al.* [49]	cross-modal	Transformer	Transformer	IU X-Ray, MIMIC-CXR	NLG metrics, CE metrics, CIDEr
Wang *et al.* [50]	cross-modal	CNN	Transformer	IU X-Ray, MIMIC-CXR	NLG metrics
Wang *et al.* [51]	cross-modal	Transformer	Transformer	MIMIC-CXR	NLG metrics
Kong *et al.* [52]	cross-modal	Transformer	Transformer	IU X-Ray, MIMIC-CXR	NLG metrics, CE metrics
Tanwani *et al.* [53]	cross-modal	CNN	Transformer	IU X-Ray	NLG metrics
Wang *et al.* [54]	cross-modal	Transformer	Transformer	IU X-Ray, MIMIC-CXR	NLG metrics, CIDEr
Yang *et al.* [55]	cross-modal	Transformer	Transformer	IU X-Ray, MIMIC-CXR	NLG metrics, CE metrics, CIDEr
Huang *et al.* [13]	cross-modal	Transformer	Transformer	IU X-Ray, MIMIC-CXR	NLG metrics, CE metrics
Wang *et al.* [12]	cross-modal	Transformer	Transformer	IU X-Ray, MIMIC-CXR	NLG metrics, CE metrics, CIDEr
Nicolson *et al.* [56]	cross-modal	CNN	LSTM	IU X-Ray, MIMIC-CXR	NLG metrics, CIDEr

vital on condition of their quality. Specifically, they contains different granular features of RRG containing key insights that guide report writing in the following aspects. First, a radiograph panoramically presents the overall health condition of a patient, which is highly associated with the consistency of the resulting report. Second, particular regions of a radiograph indicate the status of a patient's specific organs, which determines the preciseness of describing particular disease attributes in the report. Third, both global and regional features of a radiograph jointly provide necessary evidences for decision-making process for professional radiologists, especially when they write well-elaborated radiology reports. Similar to the above analysis on different features, visual-only approaches, which focus on processing radiographs, are thus proposed to facilitate RRG through extracting different granular features. In details, visual-only approaches generally consist of a visual feature extractor f_{fe} and a visual feature encoder f_{ve}, where f_{fe} is employed to encode representative features $\mathcal{H}$ from $\mathcal{I}$ by

$$\mathcal{H} = f_{fe}\left(\mathcal{I}\right) \tag{3}$$

and f_{ve} processes the resulted $\mathcal{H}$ to obtain the latent representation $\mathcal{H}^v$ through

$$\mathcal{H}^v = f_{ve}\left(\mathcal{H}\right) \tag{4}$$

Therefore, in the following texts, we categorize visual-only approaches based on the granularity of $\mathcal{H}$, including global, regional, and global-regional aggregated visual features, respectively, whose architectures are presented in Figure 2.

2.1.1 Global Visual Features

Approaches in this category focus on the global view of radiographs, where global visual feature is the most widely used feature type in existing studies. Figure 2(a) presents the visualization of approaches that adopt global visual features. According to Eq. (3), there are approaches employ

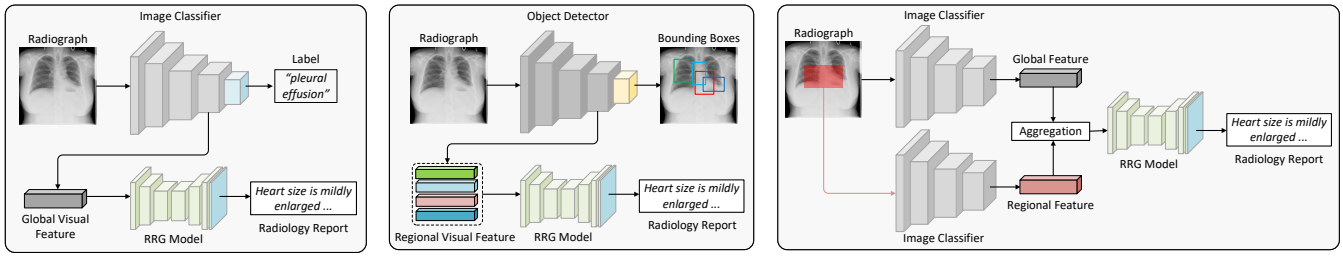


Fig. 2. The architectures of three main categories of visual-only approaches that utilize different types of visual features for the report generation process, including (a) global visual features, (b) regional visual features, and (c) global-regional aggregated features.

image classification models pre-trained on ImageNet [57] as f_{fe}, and fine-tune f_{fe} with the training objective of RRG. The global visual feature $\mathcal{H}$ is obtained from the last layer of f_{fe} based on Eq. (3), and is further processed by f_{ve} in Eq. (4) into various formations that represent the high-level semantics of radiographs, e.g., latent representation [5], [8], [15], [18], [21], [58], medical terms [7], [9], [20], [22], and graph [16], [24], [25]. Particularly, early approach [5] utilizes pre-trained VGG-Net [59] for global visual feature extraction of radiographs, and subsequent studies [8], [10], [13], [14], [16], [17], [18], [21], [22], [24], [25], [26], [27] manage to employ deeper and more complex classification models, e.g., ResNet [60] and DenseNet [61], to extract representative features from radiographs. With the development of Transformer [62] architecture in the computer vision community, vision transformer (ViT) [63] demonstrated outstanding performance on image understanding, which is also adopted by recent RRG studies [12], [43] to extract more discriminative visual feature from radiographs.

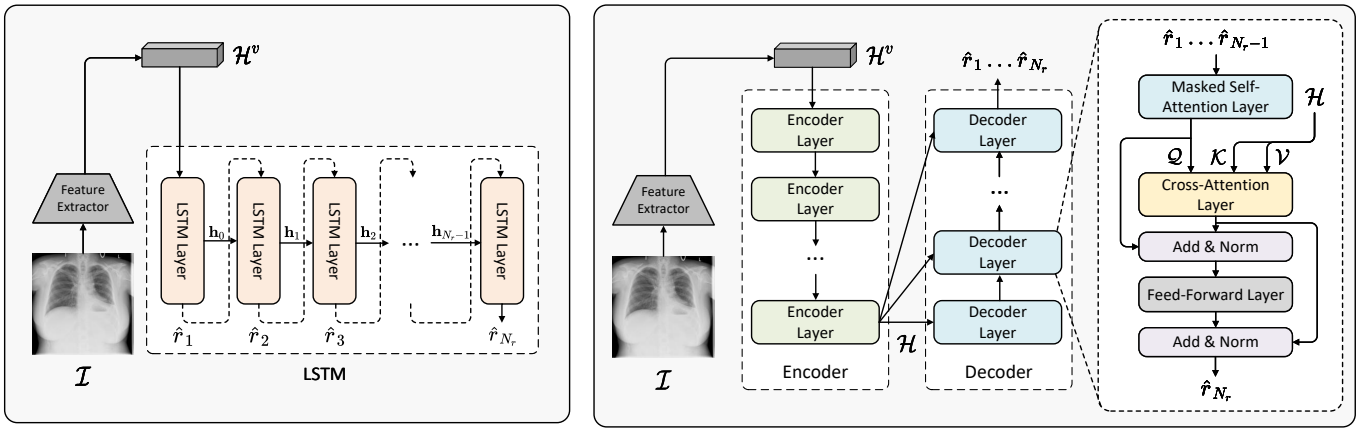
2.1.2 Regional Visual Features

Although global visual features dominate visual-only approaches, they are still ambiguous to represent specific regions of radiographs, which potentially lead to inaccurate descriptions of particular disease attributes in the generated reports. To provide more fine-grained visual information for RRG, some approaches [23], [28] extract regional features radiographs. Figure 2(b) presents the visualization of approaches that adopt regional visual features. These approaches put emphasis on local pathological lesions in radiographs, and extract features from them to provide conditional guidance to describing particular disease attributes (e.g., abnormalities). In doing so, they need to first obtain a series of anatomical regions $\{\mathcal{I}_1 \dots \mathcal{I}_{N_a}\}$ in N_a categories based on human annotations or object detection algorithms. Then, they fine-tune pre-trained object detection models as f_{fe} according to Eq. (3), and extract $\mathcal{H}$ from $\{\mathcal{I}_1 \dots \mathcal{I}_{N_a}\}$.

To facilitate RRG with regional visual features, early approaches mainly utilize hand-crafted object detection algorithms to detect regions. For example, Wang *et al.* [28] adopts a selective search algorithm [64] to generate regions for each radiograph in an unsupervised manner, and employ rules to select result regions from them. The approach then extracts features from the selected regions, and utilizes a self-attention layer to learn the internal relationships among the regions, results in relation-enhanced features for the subsequent report generation process. However, in doing so, it faces a problem in effectively locating specific organs in radiographs since the regions detected by hand-crafted object detection algorithms (e.g., selective search algorithm) usually do not precise and overlap with each other. To address the problem, there are studies that further adopt deep learning-based object detectors (e.g., Fast R-CNN [65]) to improve region detection and then achieve outstanding RRG performance on benchmark datasets. For example, Tanida *et al.* [23] leverage the Chest ImaGenome dataset [66] with 29 annotated anatomical regions to fine-tune an anatomy-based object detector that is initialized from ImageNet [57], and further use it to provide regional visual features to guide radiology report generation. To summarize, regional visual features allow RRG models to leverage fine-grained visual signals from radiographs to generate more precise texts so that ensuring them containing details of particular pathology information in the final reports.

2.1.3 Global-Regional Aggregated Visual Features

Consider that employing visual features globally or locally has demonstrated its advantages, it is thus natural to further consider join both global and regional visual features to enhance RRG, with the global features ensuring the overall consistency of the generated report, and the regional features facilitating the preciseness of produced texts to describe specific diseases. However, it is not trivial to combine them since global and regional features differ from each other in both contents and structures, which causes the feature misalignment problem. Towards solving this problem, recent studies [24] are developed to integrate global and regional features by adopting a dual-branch feature extraction pipeline to process the input radiograph, and fuse the extracted features from both branches into an integrated one, which is then used by subsequent networks to provide signals from different views (i.e., global and regional) for the report generation process. In other words, the visual feature extractor f_{fe} consists of three components for approaches using global-regional aggregated visual features, namely, a global feature extractor f^g_{fe}, a regional feature extractor f^r_{fe}, and a fusion process f_a to combine the global and regional features to get $\mathcal{H}$, whose process is illustrated in Figure 2(c). Specifically, given a radiograph $\mathcal{I}$, the global branch employs f^g_{fe} to extract the global visual feature $\mathcal{H}^g$ from $\mathcal{I}$; the regional branch utilizes f^r_{fe} to first obtain specific regions from $\mathcal{I}$, and then extract regional visual features $\mathcal{H}^r$ from the resulting regions; and f_a is used to integrate $\mathcal{H}^g$



(a) (b)

Fig. 3. The architecture of autoregressive models for textual-only approaches, including (a) long-short term memory (LSTM) and (b) Transformer.

and $\mathcal{H}^r$ into $\mathcal{H}$. Thus, the overall process is formulated by

$$\mathcal{H} = f_a\left(\mathcal{H}^g, \mathcal{H}^r\right), \quad \mathcal{H}^g = f_{fe}^g\left(\mathcal{I}\right), \quad \mathcal{H}^r = f_{fe}^r\left(\mathcal{I}\right) \tag{5}$$

where $\mathcal{H}$ is then processed by f_{ve} into $\mathcal{H}^v$, so that $\mathcal{H}^v$ is able to facilitate the report generation process with aggregated information from both global and regional views of $\mathcal{I}$. As a representative study, Li *et al.* [24] extract both global and regional features from input radiographs, and integrate them to facilitate RRG. Specifically, extracting global features in this approach employs a DenseNet-121 (a variant of DenseNet with 121 layers) as f_{fe}^g to extract $\mathcal{H}^g$ from $\mathcal{I}$. For regional features, it selects important regions in the radiograph through a masking mechanism, where features lower than a threshold in the representation of $\mathcal{I}$ are filtered out and the rest features are kept, and then uses another DenseNet-121 (serving as f_{fe}^r) to extract regional visual features $\mathcal{H}^r$. Finally, an element-wise summation operation f_a is adopted to fuse $\mathcal{H}^g$ and $\mathcal{H}^r$ and obtain the integrated feature $\mathcal{H}$ encoded by f_{ve} to facilitate report generation. Similarly, Wang *et al.* [36] consider both global and regional anatomical regions of radiographs for RRG, and propose a self-adaptive fusion gate module to dynamically aggregate global and regional visual features, so as to improve RRG. In leveraging both global and regional visual features, RRG models are able to have a comprehensive understanding of the overall information and the specific details of the radiograph, and generate better reports compared with that only enhanced by one of the feature types.

2.2 Textual-only Approaches

As the target of the RRG task, reports reveal some special characteristics of the texts associated to radiographs. First, they usually contain enriched medical information such as particular terms in describing diseases and organs, where such medical information are highly correlated with each other and indicate relationships among specific entities (e.g., diseases). Second, they are normally highly patternized, with significant structural information (e.g., formats or templates for the entire report or descriptions of diseases) to formulate the entire report. Third, they are usually long comparing to other type of texts such as normal image captions, especially when requiring to describe complicated abnormalities in the radiographs, leading to the error-prone report writing process for physicians. To consider the aforementioned properties of radiology reports, there are textual-only approaches proposed by employing different information to facilitate the report generation process. Normally, existing textual-only approaches are formulated in the following manner. Once the input radiograph $\mathcal{I}$ is processed by a visual encoder, an information extractor f_{ie} is adopted to extract information $\mathcal{K}$ from $\mathcal{H}^v$ and a text generator f_{tg} takes $\mathcal{H}^v$ and $\mathcal{K}$ to produce the final report $\widehat{\mathcal{R}}$ by

$$\widehat{\mathcal{R}} = f_{tg}\left(\mathcal{H}^v, \mathcal{K}\right), \quad \mathcal{K} = f_{ie}\left(\mathcal{H}^v\right) \tag{6}$$

Existing approaches usually adopt autoregressive models (e.g., LSTM [67] and Transformer [62]) as the f_t to produce texts (i.e., normally words) sequentially. For example, approaches [5], [14], [16], [19], [28] utilizing LSTM as f_t use the extracted visual feature as the initial hidden state of LSTM, and predict descriptive words one by one for the final report, with the details of their model architecture demonstrated in Figure 3(a). Other approaches [8], [10], [15], [18], [20], [21], [23], [24], [27], [30] use Transformer as f_t to produce the entire report through sequentially predicting the next word at each step based on multi-head attention mechanism, whose details are demonstrated in Figure 3(b). According to the aforementioned characteristics of radiology reports and the approaches proposed to address them, we categorize existing textual-only approaches into three groups, including knowledge enhancement, report structuralization, and progressive generation, whose details are illustrated as follows with elaborated descriptions.

2.2.1 *Knowledge Enhancement*

Radiology reports normally contain crucial medical information, such as medical terms and their relations, which are essential for presenting important facts in the reports. Specifically, medical terms are critical in identifying specific disease categories and attributes, as manifested in radiographs, while relations among these terms not only reveal the associations between diseases and affected organs but also provide valuable insights to them, they both

provide significant enhancement to the quality and accuracy of reports. Therefore, knowledge enhancement-based approaches [10], [15], [16], [17], [20], [22], [25], [26], [43], [43] are motivated and facilitate RRG through discovering and leveraging such crucial information for RRG, particularly through medical terms, entities and relations, as well as external help of knowledge graphs. Details of these approaches are introduced in the following texts.

Medical Terms are generally performed as one of the most commonly used knowledge types in radiology reports, which reveal detailed descriptions of particular diseases. Therefore, for RRG, medical terms are able to provide additional hints for generating descriptions of specific diseases and their corresponding attributes. Existing approaches mainly train an image tagger to extract medical terms $\{\widehat{t}_1 \dots \widehat{t}_{N_t}\}$ from radiographs, and use the terms as the information $\mathcal{K} = \{\widehat{t}_1 \dots \widehat{t}_{N_t}\}$ to enhance RRG. Therefore, the process for RRG is formulated as

$$\{\widehat{t}_1 \dots \widehat{t}_{N_t}\} = f_{ie}\left(\mathcal{H}^v\right) \tag{7}$$

and

$$\widehat{\mathcal{R}} = f_{tg}\left(\mathcal{H}^v, \{\widehat{t}_1 \dots \widehat{t}_{N_t}\}\right) \tag{8}$$

where, in order to improve the quality of extracted medical terms, f_{ie} is often trained on auto-annotated data. In doing so, people adopt existing medical term annotators, e.g., CheXpert [68] and MeSH[1], to extract N_{t^*} medical terms $\{t_1^* \dots t_{N_{t^*}}^*\}$ from the gold standard report $\mathcal{R}^*$, and use them as the input to train f_{ie}, which allows f_{ie} to extract medical terms that are more likely to appear in the report from the radiograph features and thus provide better guidance for RRG. For example, You *et al.* [10] perform RRG through disease prediction. They firstly extract visual features from the input radiograph $\mathcal{I}$ and construct a joint semantic space to align both visual and textual features, resulting in integrated global visual representation $\mathcal{H}^v$. Then, they adopt a classification layer and GRU layers as f_{ie} to predict the medical terms $\{\widehat{t}_1 \dots \widehat{t}_{N_t}\}$ based on $\mathcal{H}^v$, which results in the representation (denoted as $\mathcal{H}^t$) of $\{\widehat{t}_1 \dots \widehat{t}_{N_t}\}$. To optimize f_{ie}, this approach minimizes the negative log likelihood between model predictions $\{\widehat{t}_1 \dots \widehat{t}_{N_t}\}$ and the gold standard medical terms $\{t_1^* \dots t_{N_{t^*}}^*\}$. To predict $\{\widehat{t}_1 \dots \widehat{t}_{N_t}\}$, the approach adopts two types of labels as gold standard labels, namely, MeSH-annotated labels and context-aware labels, which are extracted by an external medical labeler named MeSH and by human annotators from the gold standard report, respectively.

However, directly predicting disease labels fails to judge the status of the predicted disease categories, which is still limited in facilitating the report generation process. Therefore, Kale *et al.* [22] aim to predict medical terms from visual features and perform a disease classification upon the estimated terms. The approach adopts an image tagger f_{ie} to classify whether the diseases are presented in the radiograph. In doing so, it employs the visual feature extractor f_{fe} (i.e., ResNet-50 [60]) in Eq. (3) to extract the visual feature $\mathcal{H}^v$ from the input radiograph $\mathcal{I}$, and utilize several fully-connected projection layers to compute the scores $\{c_1 \dots c_{N_t}\}$ of different medical terms $\{t_1^* \dots t_{N_t}^*\}$ based on $\mathcal{H}^v$. Herein, each score $c_i \in \{c_1 \dots c_{N_t}\}$ is converted into binary value (i.e., c_i equals to 0 or 1) based on a cutoff threshold T, where $c_i = 1$ and $c_i = 0$ indicate that the disease is presented or absent, respectively. The corresponding representation $\mathcal{K}$ of the medical terms is then encoded into latent representation to generate the final report $\widehat{\mathcal{R}}$. To obtain $\{t_1^* \dots t_{N_t}^*\}$, the approach fine-tunes a CNN-based classifier (i.e., ResNet) and CheXpert to annotate medical terms. Similarly, Yuan *et al.* [38] propose to predict medical terms and fuse the resulting features with visual features, so as to enhance RRG. It is worth noting that the approach considers both frontal and lateral views of the radiographs to perform RRG, where the radiographs are concatenated and used as a single radiograph for further processing. Similarly, Wang *et al.* [51] adopt predicted medical terms from the extracted visual feature of radiographs, and use the representation of the predicted terms to offer semantic assistance for the report generation process. Harzig *et al.* [39] classify whether the generated sentences are describing an abnormal case, so as to facilitate the model in generating more precise descriptive sentences. Furthermore, Nishino *et al.* [20] perform a planning-based report generation based on annotated medical terms. The approach first employs a medical term annotator (i.e., VisualCheXBERT [69]) as f_{ie} to extract medical terms $\{\widehat{t}_1 \dots \widehat{t}_{N_t}\}$ from radiographs, and adopts a content planner to process $\{\widehat{t}_1 \dots \widehat{t}_{N_t}\}$ into a series of plans, where each plan suggests the order in which the terms appear. And the approach utilizes a text generator (i.e., T5 [70]) to produce descriptive sentences according to the plans, and construct the final report according to the plan order. Then, it adopts a content planner to process $\{\widehat{t}_1 \dots \widehat{t}_{N_t}\}$ into a series of plans, which suggest the order in which the terms appear. Finally, the approach utilizes a text generator to produce descriptive sentences based on the plans, and concatenates all sentences to get the final report.

Entities and Relations are pivotal elements to characterize the associations between different components in text. In the setting of RRG, entities normally denote a series of medical terms (such as disease categories, organs, attributes, etc.), and relations usually indicate the associations among different terms (such as anatomical location, diagnostic indication, disease attributes, health status, etc.), which offer direct assistance for the report generation process. To leverage the information, existing approaches [41] formulate entities and relations by a list of tuples $\mathcal{U} = u_1 \cdots u_{N_e}$ where the n_e-th tuple $u_{n_e} = (e_{n_e,1}, r_{n_e}, e_{n_e,2})$ with $e_{n_e,1}$, $e_{n_e,2}$ denoting entities, and r_{n_e} referring to their relation. Herein, $\mathcal{U}$ is utilized as extra information $\mathcal{K}$ to facilitate the report generation process along with entities and relations. For example, Dalla Serra *et al.* [41] firstly utilize RadGraph [71] and ScispaCy [72] to extract entities and relations from the gold standard report $\mathcal{R}^*$. Herein, the entities represent the abnormalities and organs, and the relations indicate their relationships, including "*suggestive of*", "*located at*", and "*modify*". Then, the approach employs a tuple extractor f_{te} to predict the entity and relation tuples based on the extracted feature $\mathcal{H}^v$ from the input radiograph $\mathcal{I}$. Afterwards, the tuple is used as the condition of report generation, where the text decoder f_t generates the final report $\widehat{\mathcal{R}}$ based on $\mathcal{H}^v$ and u_{n_e}.

Knowledge Graph records structural relationships among

1. https://www.nlm.nih.gov/mesh/meshhome.html

different term-based information, which usually consists of nodes and edges that directly indicate the internal correlation between diseases and organs in the medical field, thus is potential to help RRG with effective guidance. However, constructing a decent knowledge graph for RRG is not a trivial task, where the settings of its nodes and edges should be reasonable and helpful to the RRG process. Existing approaches to building the knowledge graph are categorized into two groups, namely, prediction-based approaches and pre-construction-based approaches. The former approaches usually predict the probability value of nodes and edges based on extracted visual features; the latter ones pre-define the nodes and edges of knowledge graphs based on clinical rules. They usually construct the nodes to represent specific disease categories or organs, and construct the edges to refer to the correlations among diseases or associations among diseases and particular organs. Once the knowledge graph $\mathcal{G}$ is constructed, the aforementioned approaches employ a graph encoder f_{ie} to extract information $\mathcal{K}$ from $\mathcal{G}$ with the guidance of $\mathcal{H}^v$, and then utilize the text generator f_{tg} to generate the final report. In performing the aforementioned processes, Li *et al.* [15] firstly propose to perform a knowledge-enhanced RRG through three main steps, namely, encoding, retrieving, and paraphrasing. Given a radiology image $\mathcal{I}$, the encoding step of the approach extracts the visual feature $\mathcal{H}^v$ from $\mathcal{I}$, and encodes $\mathcal{H}^v$ into an abnormality graph $\widehat{\mathcal{G}}$. The node and edges in $\widehat{\mathcal{G}}$ are compared with the gold standard graph $\mathcal{G}^*$, where the binary cross-entropy loss is computed and used to optimize the model. To obtain $\mathcal{G}^*$, the approach defines clinical abnormalities stemming from thoracic organs as nodes of $\mathcal{G}^*$, and uses the co-occurrence of abnormalities as edges. Then, the retrieving step obtains report templates based on the predicted knowledge graph, and the paraphrasing step refines the retrieved templates with enriched descriptions so as to generate the final report $\widehat{\mathcal{R}}$ accordingly.

However, approaches based on predicted knowledge graphs present a limitation where they are only able to predict a part of the entire graph, which makes it hard for them to get fully enhanced by knowledge graphs to improve RRG. To address the limitation, there are studies that pre-construct the knowledge graphs instead of predicting them, so as to provide high-quality knowledge graphs to assist the report generation process. In doing so, Zhang *et al.* [16] manage to utilize a pre-constructed knowledge graph to assist the report generation process. The approach firstly pre-trains an image classification model f_{fe} on CheXpert [68] to classify a series of medical findings corresponding to the nodes of a pre-defined knowledge graph $\mathcal{G}$, where f_{fe} is used to extract the visual feature $\mathcal{H}^v$ from the input radiograph $\mathcal{I}$. Then, it conducts the attention mechanism over $\mathcal{H}^v$ to initialize the knowledge graph $\mathcal{G}$, and employs graph convolutional networks (GCN) [73] to propagate information across nodes and edges of $\mathcal{G}$, where $\mathcal{G}$ is further optimized through a multi-label classification process. Finally, $\mathcal{G}$ is encoded into graph embedding $\mathcal{H}^G$ for the text decoder, which generates the final report $\widehat{\mathcal{R}}$ with the assistance of $\mathcal{H}^G$. To define $\mathcal{G}$, the approach extracts 20 categories of keywords from the radiology reports as the nodes of $\mathcal{G}$, and utilizes its edges to represent the relationship between diseases and particular organs. The knowledge graph constructed by Zhang *et al.* is widely exploited in the follow-up studies [17], [43]. Specifically, Liu *et al.* [17] further utilize medical terms as prior knowledge, and adopt related reports and pre-constructed knowledge graph as posterior knowledge to enhance RRG. Kale *et al.* [25] construct a knowledge graph for X-ray images to augment a pre-trained visual-language BART [74], so as to improve RRG. Particularly, the approach constructs $\mathcal{G}$ based on extracted term and relation tuples from radiology reports following the approach of Kale *et al.* [75]. It defines eight logical relations based on a series of medical characteristics, including anatomy, entities type, anatomical locations and properties, findings, etc., as the edges of $\mathcal{G}$, and integrates several categories based on Radlex entities[2] as the nodes of $\mathcal{G}$. Recent studies consider the deficiency of pre-constructed knowledge graphs in conventional graph-based approaches, where the pre-defined scope of knowledge hurts the effectiveness of these approaches. Therefore, Li *et al.* [43] propose to extend the pre-constructed knowledge graph, so as to facilitate the RRG process with both pre-defined and additional knowledge. The approach regards the pre-constructed knowledge graph $\mathcal{G}$ [16] as general knowledge, and performs a dynamic graph construction process upon $\mathcal{G}$ to introduce report-specific knowledge into $\mathcal{G}$. In doing so, the approach firstly retrieves top-K the most similar reports based on the similarity between the extracted visual feature $\mathcal{H}^v$ and report feature $\mathcal{H}^r$. Then, it extracts anatomy and observation entities by Stanza [76], and utilizes the resulting entities to obtain report-specific knowledge $\mathcal{K}$ through RadGraph [71]. To extend the pre-constructed knowledge graph $\mathcal{G}$, the approach employs relations in RadGraph as new edges of the extended graph $\widehat{\mathcal{G}}$, and adds $\mathcal{K}$ as additional nodes of $\widehat{\mathcal{G}}$. Finally, $\widehat{\mathcal{G}}$ is capable of facilitating the report generation process with both general and report-specific knowledge, resulting in the generated report $\widehat{\mathcal{R}}$. Similarly, Yan *et al.* [42] enrich the constructed knowledge graph with abnormality nodes and attribute nodes, and propose an automatic approach to build the knowledge graph based on annotations, radiology reports, and RadLex radiology lexicon.

In leveraging different types of information, RRG models are capable of generating better reports with more focus on particular information, e.g., specific terms and relationships between terms. Among this information, medical terms provide term-level information for report generation, entities and relations further indicate the relationship between every two terms, and the knowledge graph offers a holistic view of different terms (nodes) and their relationships (edges), where more information are provided for the report generation process from medical terms to knowledge graph, thereby generating reports of higher elaboration.

2.2.2 Report Structuralization

One of the most remarkable characteristics of radiology reports is that reports are usually highly patternized. Specifically, physicians normally write descriptions following particular sentence structures of diseases and organs, and follow certain templates to construct the entire report. This

2. http://radlex.org/

observation indicates that radiology reports consist of significant structure information, which presents the potential to enhance the report generation process by filling in enriched details in describing diseases based on the confirmed structures. Therefore, some studies [15], [22], [27], [41] propose to process radiology reports into particular patterns based on pre-defined rules, where the report generation process is then guided by these patterns so as to enrich the description details accordingly. Particularly, these studies adopt two main categories of patterns to facilitate RRG, namely, report template and report clustering. Details of each specific category are illustrated in the following texts.

Report Templates provide guidance for RRG that is neglected by entities and relations. Normally, the approach based on report template firstly determines the overall structure of the generated report and then allows RRG approaches to fill in descriptive contents into these templates according to the radiology image. In doing so, existing approaches [15], [22] mainly employ descriptive sentences of the same abnormality as report templates for RRG. They perform RRG in a two-stage manner, including template construction and report generation processes. The template construction process obtains the report template by extracting patterns from all radiology reports in the entire training set, and the report generation process fills descriptive details into the report template, thereby producing the final report.

Specifically, Li *et al.* [15] utilize similar descriptive sentences of different abnormalities as templates to assist the report generation process. For each specific abnormality, the approach collects all sentences that describe the abnormality in the training corpus, where the sentences with the same meaning are grouped, and the most frequent one is used as the template. Given the input radiograph $\mathcal{I}$, the report template is able to serve as guidance for generating descriptions for each particular abnormality, thereby constructing the entire report $\widehat{\mathcal{R}}$ according to $\mathcal{I}$. Later, Kale *et al.* [22] construct a more fine-grained report template, with specific labels of described organs as templates for the contents of sentences in the final report. For example, each template is represented by a particular term, indicating the abnormality region that the sentence is describing, e.g., "*lung1*" refers to the first template type to describe the lung area. From another aspect, while existing studies adopt pre-defined report templates for RRG, Chen *et al.* [8] propose relational memory networks to automatically store the report template information and use it to guide the report generation process. Particularly, the relational memory networks are conducted as a matrix $\mathcal{M}_t$, where $\mathcal{M}_t$ at the t-th generation step is updated to $\mathcal{M}_{t+1}$ for the next step during the generation processes of reports. Also, the approach adopts a memory-conditional layer normalization mechanism to retrieve the template information from the memory matrix and incorporate it into the report generation process, so as to facilitate RRG with the recorded templates. Similarly, Yu *et al.* [48] adopt a memory bank to preserve key information for radiology report generation, and employ reinforcement learning to improve the model to generate clinically accurate reports that contain critical information, such as the presence status of different diseases.

Report Clustering considers features shared by similar reports (e.g., relevant phrases or identical expressions) that are associated with the same or similar radiographs, which provides a reference for physicians when writing reports even if an input radiograph is previously unseen. In doing so, the existing approach clusters all radiology reports in the training corpus into a series of categories, where reports in the same category are considered to be semantically close to each other. Thus, the report clustering process provides additional supervision signals for the optimization of the RRG pipeline, thereby facilitating the report generation process.

Specifically, Yan *et al.* [27] propose a weakly supervised contrastive learning with clustered reports to facilitate RRG. In doing so, the approach first projects all radiology reports in the training corpus onto a semantic space through CheXBERT [77], and conducts a K-means clustering algorithm upon the report representations. Considering that reports in the same category are semantically similar to each other, the approach optimizes the model to generate reports that are close to the semantically close ones through contrastive learning. Then, the similarity of positive pairs is maximized and the one of negative pairs is minimized in a way of contrastive learning, which finally assists the overall pipeline to generate an image-text aligned report $\widehat{\mathcal{R}}$.

In performing report structuralization for RRG, RRG models are able to utilize the pattern information shared by different reports to assist the report generation process. Specifically, report templates are carefully constructed according to the radiology reports, which requires an additional manual selection process to improve RRG models. Report clustering acts as an automatic solution to extract pattern information from reports and provides heuristics upon automatic report structuralization.

2.2.3 Progressive Report Generation

Although existing textual-only approaches have adopted a series of textual characteristics for RRG, the one-step generation process of the entire report is still challenging. Therefore, progressive report generation is promoted, which decomposes the RRG task into multiple sub-tasks, so as to alleviate the difficulty of generating full radiology reports. In doing so, existing approaches [9], [37] manage to perform RRG in a two-stage manner. Given a radiology image $\mathcal{I}$, they firstly generate an intermediate report $\widehat{\mathcal{Z}}$ (e.g., high-level summarization of the entire report) according to $\mathcal{I}$, and produce the final report $\widehat{\mathcal{R}}$ according to $\widehat{\mathcal{Z}}$ progressively.

Specifically, Xue *et al.* [37] manage to generate radiology reports in a sentence-by-sentence paradigm, where the study adopts the previously produced sentence as the condition to generate the next one, thereby ensuring the textual coherence in the resulted report. Compared to sentence-level progressive generation, Nooralahzadeh *et al.* [9] propose a consecutive generation framework to perform a two-stage report generation process. The approach firstly processes the gold standard report $\mathcal{R}^*$ to sequentially extract the disease words, negation or uncertainty of each word, and disease attributes from $\mathcal{R}^*$ based on dependency graph parsing, which results in summarization with respect to disease categories and attributes Given the input radiograph $\mathcal{I}$, the approach utilizes a language model (i.e., BART [74]) to firstly produce summarization, and then generates the final

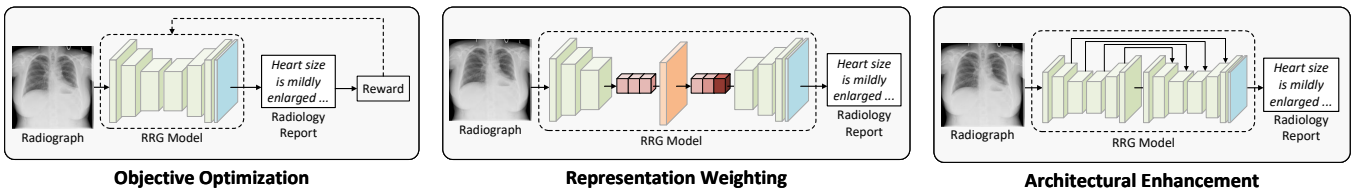


Fig. 4. The architectures of three main categories of cross-modal approaches that enhance the cross-modal alignment for RRG from different perspectives, including (a) objective optimization, (b) representation weighting, and (c) architecture enhancement.

report $\widehat{\mathcal{R}}$ according to the summarization. By decomposing the entire report generation process into multiple steps, RRG models are able to perform RRG in a coarse-to-fine manner, thereby generating more precise report than the conventional one-step generation paradigm.

2.3 Cross-modal Approaches

Owing to the cross-modal nature of RRG, it is straightforward to consider leveraging the relationship between the two modalities, i.e., vision (radiographs) and text (reports), where such relationship actually establishes the connection between the characteristics from both sides. Yet, utilizing the radiograph-report connection is not a trivial task, where the cross-modal mapping has its own characteristics in the following aspects. First, human preferences play a pivotal role in choosing different ways to employ such connection, where physicians usually construct reports based on their own knowledge such as clinical experiences. Second, radiograph regions usually receive imbalanced attentions in report writing, so that leads to different descriptive sentences with variant levels of details in the final report. Therefore to consider these characteristics, cross-modal approaches are motivated to optimize or enhance the alignment between radiographs and reports to facilitate RRG. Specifically, some approaches [6], [7], [11], [28], [78] aim to improve the optimization objectives of the RRG process from both modalities by requiring additional supervision signals (e.g., medical term annotations). Meanwhile, other studies [5], [8], [18], [19], [21], [29] conduct representation weighting for different modalities in order to balance the contribution from either visual or textual guidance, thus enhance fine-grained alignment between radiographs and reports to facilitate RRG. Later, based on the aforementioned work, recent approaches [12], [13] enhance the cross-modal encoding and decoding from the perspective of model architecture, and integrate additional information to further improve the encoder and decoder for RRG. Therefore, we categorize cross-modal approaches into three groups, emphasizing on objective optimization, representation weighting, and architecture enhancement. These approaches in three categories are illustrated in the following texts with detailed information.

2.3.1 Objective Optimization

Objective optimization-based approaches [5], [7], [8], [11], [12], [13], [14], [18], [19], [21], [27], [28], [29], [78] enhance the cross-modal alignment between radiographs and reports by improving the training objectives, so as to facilitate RRG. The architecture of this approach is shown in Figure 4(a). We classify existing approaches into four main categories to optimize the training objectives, including reinforcement learning, curriculum learning, self-boosting, and pre-training whose details are illustrated in the following texts.

Reinforcement Learning provides an alternative learning objective rather than the standard supervised learning that typically employs a cross-entropy loss function for optimization (as shown in Eq. (2)), which is prone to exposure bias [79] and does not always align with evaluation metrics [79]. In contrast, reinforcement learning is able to directly optimize performance based on different metrics by using these metrics as rewards to update model parameters. The key elements of reinforcement learning include an agent, environment, action, policy, and reward. Several approaches for RRG [6], [7], [11], [14] treat the generation model as the agent interacting with an external environment (i.e., features in different modalities), where the model parameters θ, which is represented as policy p_θ, dictate the actions taken, such as predicting the entire report or the next token. These approaches focus on designing various evaluation metrics, such as BLEU [80], METEOR [81], and ROUGE [82], and use the combination of them as rewards r to optimize the generation model. The primary objective of reinforcement learning is to maximize the expected reward for a prediction $\widehat{y}$ made by the model, which is expressed as

$$L(\theta) = \mathbb{E}_{\widehat{y}\sim p_\theta}\left[r(\widehat{y})\right] \tag{9}$$

The gradient of $L(\theta)$ with respect to θ is computed using the REINFORCE algorithm [83] through

$$\nabla_\theta L(\theta) = -\mathbb{E}_{\widehat{y}\sim p_\theta}\left[r(\widehat{y})\nabla_\theta \log p_\theta(\widehat{y})\right] \tag{10}$$

This method ensures that actions resulting in a higher reward $r(\widehat{y})$ are more likely to be repeated in the learning process, guiding the model towards optimal performance.

Li *et al.* [6] adopt a combination of retrieval policy module and generation policy module to optimize RRG, and design a sentence-level and word-level reward for optimization. Given the input radiograph $\mathcal{I}$, it firstly employs a CNN-LSTM architecture to generate the hidden topic states $\mathcal{H}_t^r$ at t step. Then, a one-layer perceptron predicts the probability distribution over actions of generation or retrieval based on $\mathcal{H}_t^r$, where the module with higher probability is activated to generate the final report $\widehat{\mathcal{R}}$ or retrieve $\widehat{\mathcal{R}}$ from collected template databases. Herein, the approach sets a threshold to filter out infrequent sentences, and obtains the template databases according to the frequency of sentences in radiology reports. Besides, the approach designs a sentence-level reward r^s and word-level reward r^w based on CIDEr [84] following Eq. (9) and Eq. (10), so as to optimize the retrieval and generation policy modules,

respectively. Herein, r^s and r^w are computed through delta CIDEr scores upon the entire reports and generated words, respectively. Jing *et al.* [14] further refine the rewards of Li *et al.* [6], and proposes rewards of normal and abnormal sentences in a generated report to facilitate the RRG process. Given the input radiograph $\mathcal{I}$, the approach firstly generates a report $\widehat{\mathcal{R}}$ following the standard RRG process. Then, it decomposes $\widehat{\mathcal{R}}$ into a series of abnormal and normal sentences, denoting as $S^{ab} = \{s_1^{ab} \dots s_{N_ab}^{ab}\}$ and $S^{nr} = \{s_1^{nr} \dots s_{N_nr}^{nr}\}$ with N_{ab} and N_{nr} referring to the number of abnormal and normal sentences, respectively. The rewards r^{ab} and r^{nr} to generate abnormal and normal sentences in $\widehat{\mathcal{R}}$ are computed according to BLEU-4 score [80] through comparing S^{ab} and S^{nr} with the corresponding sentences S^{*ab} and S^{*nr} in the gold standard radiology report $\mathcal{R}^*$, respectively.

Previous reinforcement learning-based studies mainly consider report-level enhancement, where such rewards have limited effectiveness in describing specific diseases. Therefore, Nishino *et al.* [7] propose to improve the generated reports with a token-level reward, which adopts a clinical reconstruction score (CRS) and a data augmentation approach for reinforcement learning to address the data imbalance difficulties in RRG. In doing so, the approach adopts a medical labels reconstructor (i.e., BERT [85]) following the standard generation process of the final report $\widehat{\mathcal{R}}$, so as to reconstruct the finding labels from $\widehat{\mathcal{R}}$. Then, CRS is computed as the F1 score between the generated and the input finding labels. Besides, the approach employs an additional ROUGE-L score [82] to design the reward r in Eq. (9), where r is written as $r(\widehat{\mathcal{R}}) = \lambda f_{rouge}(\widehat{\mathcal{R}}, \mathcal{R}^*) + (1-\lambda) f_{crs}(\widehat{\mathcal{R}})$ with $f_{rouge}(\cdot)$ and $f_{crs}(\cdot)$ referring to the processes of computing ROUGE-L and CRS scores, respectively.

Curriculum Learning [86] enables neural networks to gradually proceed from easy samples to more complex ones in training, which is also adopted by the recent study [78] to gradually enhance the radiograph-report alignment for RRG. These approaches normally require a well-designed metric to measure the difficulty of the currently underlying task and adjust the training samples accordingly. However, it is challenging for the RRG task to directly apply the standard curriculum learning paradigm, which measures task difficulties solely relying on a single modality, whereas RRG is a multiple task where both radiographs and reports play essential roles in identifying the difficulty. Therefore, dedicated curriculum learning is expected for RRG.

To address this challenge, Liu *et al.* [78] propose a competence-based multi-modal curriculum learning approach with multiple difficulty metrics for RRG. The proposed approach measures the difficulty d of instances based on their difficulties in capturing and describing the abnormalities from radiographs and radiology reports, which consist of visual and textual difficulty, respectively. For visual difficulty, the approach considers a radiograph as a more difficult training instance if it contains more abnormalities. Specifically, the approach adopts a fine-tuned vision backbone model (i.e., ResNet [60]) on CheXpert dataset [68], which extracts features from all normal radiographs in the training set and computes their average feature $\mathcal{H}^{v,n}$. Given an input radiograph $\mathcal{I}$, the visual difficulty d^v is defined as the average cosine similarity between the extracted feature $\mathcal{H}^v$ of $\mathcal{I}$ and $\mathcal{H}^{v,n}$, where a higher similarity indicates $\mathcal{I}$ is an easier instance. For the textual difficulty, the approach adopts the number of abnormal sentences in a report to define its difficulty d^t. Following Jing *et al.* [5], sentences without specific words (e.g. "*no*", "*normal*", "*clear*", and "*stable*") are considered to be abnormal sentences, while the others represent normal ones. For model confidence, the approach also introduces two model-based metrics from perspectives of radiographs and reports, denoting as d_m^v and d_m^t, respectively, where d_m^v refers to the entropy value of the predicted probability distribution by the vision backbone model, and d_m^v denotes the negative log-likelihood loss values following [87], [88]. To optimize the model, the approach builds curriculum learning process based on Plananios *et al.* [78], which adjusts the applied d based on perplexity (PPL).

Self-boosting aims to boost the optimization of different components in the overall pipeline. In doing so, the overall pipeline is usually conducted in multiple branches with different training objectives, e.g., radiograph-report matching and report generation, which are used to optimize the entire RRG pipeline in an alternative manner. Herein, when a specific branch is being optimized, the model parameters of other branches are fixed. Then, the updated branch is enhanced by other branches, which eventually leads to better optimization of the overall RRG pipeline.

Specifically, Wang *et al.* [28] propose a dual-branch framework to enhance RRG in a self-boosting manner. In doing so, the overall framework of the approach is based on two branches, i.e., a report generation branch and an auxiliary image-text matching branch, where both branches facilitate each other with their internal interactions. The report generation branch supplies hard negative samples for the image-text matching branch so as to force the latter to perform better feature learning. The image-text matching branch provides visual-textual aligned features for the report generation branch, which computes a triplet loss based on the cosine similarity of radiograph-report pairs and hard negative samples following the standard process of VSE++ [89]. In addition, both branches are optimized alternatively in a way that is similar to the training paradigm of generative adversarial networks (GAN) [90], where the report generation branch focuses on producing the final radiology reports and the image-text matching branch aims to differentiate the generated and gold standard reports.

Pre-training is another optimization technique that widely used in representation learning. Since most existing RRG approaches tend to train their models from scratch, they often have difficulties in fully optimizing their models, especially the ones with complicated network architecture. Therefore, some studies adopt pre-trained model weights to provide better initialization for RRG. For example, Nicolson *et al.* [56] leverage off-the-shelf pre-trained model weights on the medical domain to enhance radiograph encoding and report generation, where they adopt pre-trained vision transformer (ViT) [63] on ImageNet [57] and PubMedBERT [91] to initialize the visual encoder and text decoder, respectively, so as to transfer the learned knowledge in pre-training to RRG.

In improving the optimization objectives, RRG models are capable of producing more elaborated reports with improved optimization. Specifically, reinforcement learn-

ing and curriculum learning obtain improved performance, which requires carefully designed reward or difficulty metrics for such improvement. Otherwise, self-boosting serves as a novel approach, which is able to achieve better optimization with dedicated model design. Although studies that adopt pre-training for RRG are limited, these approaches provide enlightening ideas to perform RRG better.

2.3.2 Representation Weighting

Representation weighting-based approaches [5], [8], [12], [13], [17], [18], [19], [21], [29] aim to enhance the cross-modal alignment through weighting the representations of different modalities for RRG, where the architecture of the approaches is illustrated in Figure 4(b). Specifically, during the generation process, these approaches compute the visual representation $\mathcal{H}^v$ of the radiograph $\mathcal{I}$ and textual representation $\mathcal{H}^t$ of the generated reports, and then calculate the weights for multi-modal representations, which are applied to $\mathcal{H}^v$ and $\mathcal{H}^t$ so that information from different modalities is aligned by the weighted sum of visual and textual representations. This process allows RRG models to highlight important information in different modalities for text generation and leverage it accordingly to produce accurate reports. Generally, representation weighting approaches are categorized into two groups: the first uses attention mechanisms, and the second utilizes memory networks. Details of the two groups of approaches are illustrated as follows.

Attention Mechanism [62], [92] serves as a powerful representation weighting approach, which is firstly introduced for machine translation. Owing to its effectiveness in identifying important contexts in the source language and leveraging them to build a bridge of translation between the source and target languages, the attention mechanism is generalized to many generation tasks, including RRG, to translate source data (e.g., radiology) into the target one (e.g., report) [5], [93]. Details of existing attention-based approaches for RRG are illustrated in the following texts.

Jing *et al.* [5] propose a co-attention mechanism for RRG to weight visual representation of radiographs with high-level semantics from medical terms. In doing so, the approach first employs a visual feature extractor (i.e., VGG-19 [59]) to obtain the visual representation $\mathcal{H}^v$ from the input radiograph $\mathcal{I}$, and computes the semantic representation $\mathcal{H}^s$ of medical terms based on $\mathcal{H}^v$. Then, it adopts the co-attention mechanism to incorporate $\mathcal{H}^v$ and $\mathcal{H}^s$ by applying attention weights to them, where the attention weights are modeled by one-layer perceptrons. Finally, the weighted representation is sent into a hierarchical LSTM to produce the generated report $\widehat{\mathcal{R}}$. Furthermore, some studies introduce additional information (e.g., clinical information) into the report generation process through an improved attention mechanism. In doing so, Liu *et al.* [78] propose multi-modality semantic attention to establish a correlation between regional image features, semantic representations of key findings, and clinical information. The approach first extracts the regional visual feature $\mathcal{H}^v$ from the input radiograph $\mathcal{I}$, and obtains the clinical feature $\mathcal{F}^c$ based on the disease, age, and sex of the patient. Then, it adopts the multi-modality semantic attention to weight the hidden state $\mathcal{H}^t$ at step t by applying the corresponding attention weights of $\mathcal{H}^v$ and $\mathcal{F}^c$ upon them. Therefore, both regional image feature and clinical information are incorporated into the text decoding process of LSTM, which is able to generate the final report $\widehat{\mathcal{R}}$ that is aligned with $\mathcal{H}^v$ and $\mathcal{F}^c$.

Recent approaches [17], [29] improve the multi-head attention mechanism in Transformer [62] to facilitate RRG. Liu *et al.* [29] propose an aggregate attention mechanism to incorporate normal radiographs for RRG, and adopts a contrastive attention mechanism to distinguish abnormalities of radiographs from normal ones, where both attention mechanisms are built based on multi-head self-attention. In doing so, the approach firstly collects a normality pool $\mathcal{P} = \{\mathcal{H}^v_1 \dots \mathcal{H}^v_{N_p}\}$ with 1,000 visual features of normal radiographs from the training set. Given the input radiograph $\mathcal{I}$, it computes the similarity between the extracted representation $\mathcal{H}^v$ of $\mathcal{I}$ and all elements in the normality pool, where weights of the most similar instance are applied upon $\mathcal{H}^v$ resulting in $\mathcal{H}^{v,n}$. Then, the representation $\mathcal{H}^{v,a}$ of abnormal radiograph is computed through subtracting the representation $\mathcal{H}^v$ of normal image, denoting as $\mathcal{H}^{v,a} = \|\mathcal{H}^v - \mathcal{H}^{v,n}\|^2_2$. Finally, the contrastive attention mechanism updates $\mathcal{P}$ and further weighs $\mathcal{H}^v$ based on $\mathcal{H}^{v,a}$, where $\mathcal{H}^v$ is utilized for the generation process of the final report $\widehat{\mathcal{R}}$. Similarly, Yan *et al.* [49] propose a variational auto-encoder (VAE) based prior encoder to extract prior features from radiographs, and further utilize a prior guided attention layer to weight the prior features and visual features for improving RRG.

Memory Networks [98] play as an auxiliary component firstly proposed to leverage external information for question answering, where a list of memory vectors is used to store different external information, and the vectors are retrieved and weighted based on their importance in answering the questions. Motivated by this idea, there are studies that apply memory networks to RRG, where the mechanism of memory networks performs the radiograph-report matching process, which is able to significantly facilitate the cross-modal alignment for the RRG task. Specifically, memory networks are conducted via memory matrix $\mathcal{M} = \{\mathbf{m}_1 \dots \mathbf{m}_{N_m}\}$, which stores N_m memory vectors to interact with features in different modalities. Existing approaches to adopt cross-modal memory networks for RRG generally contain the following steps. Given the input radiograph $\mathcal{I}$ and the corresponding report $\mathcal{R}^*$, the visual representation $\mathcal{H}^v$ and textual representation $\mathcal{H}^t$ are firstly computed from $\mathcal{I}$ and $\mathcal{R}^*$. Then, $\mathcal{H}^v$ and $\mathcal{H}^t$ interact with $\mathcal{M}$ based on two processes, namely memory querying and memory responding, respectively. Memory querying uses $\mathcal{H}^v$ and $\mathcal{H}^t$ to retrieve the top-$\mathcal{K}$ most related vectors from $\{\mathbf{m}_1 \dots \mathbf{m}_{N_m}\}$ and obtain their corresponding weights $\{\omega_1 \dots \omega_{N_m}\}$. Memory responding applies the weights $\{\omega_1 \dots \omega_{N_m}\}$ to the memory vectors $\{\mathbf{m}_1 \dots \mathbf{m}_{N_m}\}$, which results in the responded vector denoted by $\mathbf{r}$.

Owing to the lack of annotated alignment to map cross-modal information for RRG, Qin *et al.* [21] further enhance the cross-modal alignment based on Chen *et al.* [18] with carefully designed reinforcement learning rewards. It leverages the natural language generation (NLG) metric to guide the cross-modal aligning process. Given the input radiograph $\mathcal{I}$, the approach follows the standard RRG process in Eq. (1) to generate $\widehat{\mathcal{R}}$, and employs the memory networks

TABLE 2
The statistics of all radiology datasets with respect to category, the numbers of images, reports, and patients of raw data. The number of instances (i.e., radiograph-report pairs) in training, validation, and test sets and the average word-based length of the report in each dataset are also reported. "†" marks the datasets that randomly split the training, validation, and test sets following the ratio of 7:1:2.

		Raw Data			Experiment Data			
Dataset	**Category**	Image	Report	Patient	Train	Val	Test	Avg. Len.
CX-CHR† [6]	Benchmark	45,598	45,598	33,236	31,919	4,559	9,120	-
IU X-Ray [94]	Benchmark	7,470	3,955	3,867	5,229	747	1494	36.0
MIMIC-CXR [95]	Benchmark	473,075	377,110	270,790	65,379	2,130	3,858	57.5
MIMIC-CXR-JPG [96]	Benchmark	377,110	227.835	65,379	222,758	1,808	3,269	57.5
MIMIC-ABN [97]	Benchmark	38,551	38,551	-	26,946	3,801	7,804	-
COV-CTR† [24]	Other	728	728	728	510	72	146	-
JLiverCT [20]	Other	882	1,083	60	882	127	74	54.6
PEIR Gross [5]	Other	7,442	7,442	-	-	-	-	12.0

to interact with features in different modalities. To optimize the overall pipeline, the approach adopts the standard reinforcement learning process in Eq. (9) and (10), where the reward r is the improvement on different evaluation metrics (i.e., BLEU [80]) when generating $\widehat{\mathcal{R}}$. Then, the model is trained to maximize the expected reward so as to enhance the alignment between radiographs and reports.

In leveraging representation weighting, RRG models are able to perform the report generation process better with enhanced cross-modal alignment between radiographs and reports. The attention mechanism facilitates RRG by modifying the representation with different weights, while the memory networks serve as another form of attentions, which utilizes memory vectors to compute the weights to balance the contributions of different representations.

2.3.3 Architecture Enhancement

Architecture enhancement boosts the radiograph-report alignment for RRG by modifying the architectural design of neural networks, where the architecture of the approaches is shown in Figure 4(c). In doing so, some studies [12], [13], [46] introduce the task-specific features of radiographs and reports into foundation model architecture such as Transformer [62], while the recent study [56] focuses on offering better initialization for particular model components (e.g., visual encoder or text decoder) with pre-trained model weights. Specifically, You *et al.* [46] align the visual regions of radiographs with specific medical terms using a hierarchical Transformer architecture. The approach firstly extracts visual features from the input radiograph and term features from medical terms and next concatenates them, where a multi-head self-attention layer is applied to process them, so as to learn the correlation between radiographs and medical terms. Then, the outputs of different attention layers are sent to a Transformer decoder, and are incorporated through a cross-attention layer with additional learnable parameters that modify the scales, so as to facilitate the report generation process with multiple granular visual-textual alignment. Existing approaches mainly adopt the standard Transformer architecture for RRG. For example, Hou *et al.* [45] combine a pre-trained DenseNet-121 [61] with a standard Transformer decoder for RRG. Huang *et al.* [13] propose a knowledge-injected U-Transformer architecture for RRG. The overall architecture consists of two standard Transformer, which contains skip connection between layers of the encoder and decoder to share their internal features. Besides, extra knowledge is incorporated into each Transformer to enhance the radiograph encoding and report generation processes. While improving the architecture Transformer is demonstrated to be superior effectiveness, other studies emphasize improving the input of Transformer-based models. For example, Wang *et al.* [12] add additional expert tokens into the input of Transformer-based visual encoder for RRG, where these expert tokens learn the correlations with other visual features and serve as conditions for report generation. Recently, there are studies that propose multi-task learning model to further improve report generation. In doing so, Tanwani *et al.* [53] propose a dedicated framework architecture to perform both medical visual question answering (VQA) and report generation, where the encoder aligns radiographs with text descriptions in reports through contrastive learning, and the decoder predicts the report (or answer for VQA) enhanced by related report retrieved from FAISS library[3]. In enhancing the model architecture, RRG models are able to generate elaborated reports with dedicated network architecture.

3 Datasets

To conduct clinical research with deep learning-based approaches, some studies [6], [24], [68], [94], [95], [96], [97], [99], [100] manage to construct datasets so as to push forward the development of AI-assisted applications, e.g., mutli-label chest X-ray classification [73], [101], [102], [103], [104], medical image segmentation [105], [106], [107], [108], [109], [110], [111], [112], doctor-patient communications [113], [114], [115], [116], [117], [118], as well as RRG [5], [8], [12], [13], [15], [16], [18], [21], [22], [25], and many others. As for radiograph-driven research, recent studies construct datasets by collecting chest X-ray images and the corresponding medical records (e.g., radiology reports and medical terminologies) from the hospital, and preprocess the collected data with necessary anonymization such as patient de-identification, so that ensure the quality of these datasets and keep their quantity to meet the requirements of research activities. We classify prevailing datasets for RRG studies into two main categories, namely

3. https://github.com/facebookresearch/faiss

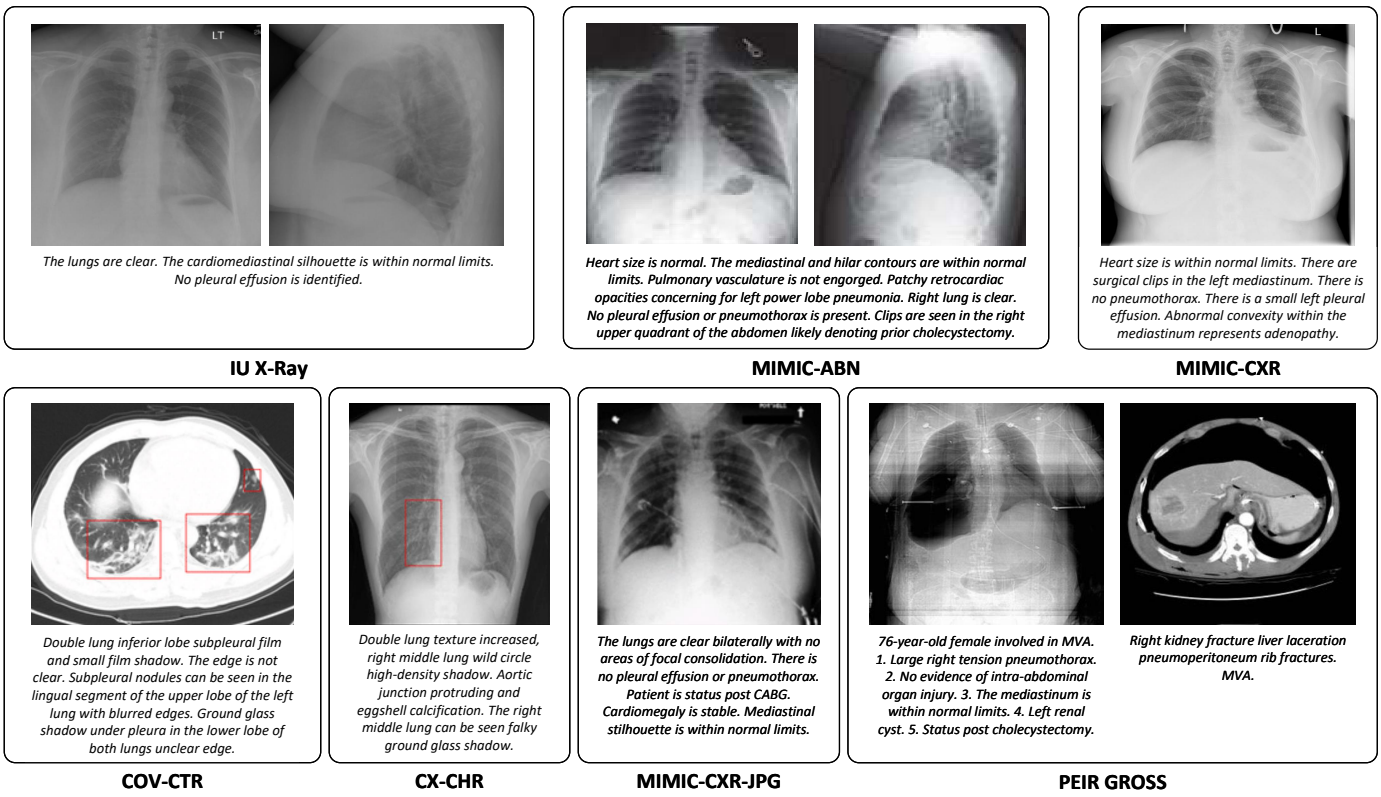


Fig. 5. Examples of radiographs and their corresponding reports in IU X-Ray [94], MIMIC-CXR [95], MIMIC-ABN [97], MIMIC-CXR-JPG [96], CX-CHR [6], COV-CTR [24], and PEIR Gross [5], where the radiology reports of CX-CHR and COV-CTR are translated from Chinese into English. The red boxes in radiographs from CX-CHR and COV-CTR are annotated by physicians to highlight the attentive regions that the reports describe.

benchmark datasets and other datasets, respectively. Figure 5 demonstrates some examples of radiographs and their corresponding reports[4], and Table 2 reports the statistics of all datasets with respect to the main characteristics, including domain, image number, report number, the number of radiograph-report pairs, patient number, and average length of radiology reports. Details of the aforementioned datasets in different categories are illustrated in the following texts.

3.1 Benchmark Datasets

Benchmark datasets [6], [94], [95] are conventionally used by most existing RRG studies to evaluate their model performance and compare it with others. Normally, these datasets contain massive amounts of image-text pairs of radiographs and radiology reports, which is capable of representing the RRG task in terms of its challenging scenarios and ideal expected results. Moreover, they are able to supply sufficient data resources for existing RRG studies to perform experiments on their approaches, as well as different analyses. Existing benchmark datasets mainly consist of **CX-CHR** [6], **IU X-Ray** [94], and **MIMIC-CXR** [95], with details of each specific dataset illustrated as follows.

CX-CHR [6] is a benchmark dataset that consists of chest X-Ray images with Chinese reports. The dataset is collected from a professional medical institution for health checking in China, along with 35,609 patients, 45,598 radiographs, and the corresponding Chinese reports. Each report contains the conventional findings, and impression sections, along with an additional complains section to reflect syndromes of the corresponding patient from different perspectives. Particularly, each radiograph is annotated with bounding boxes (see the example in Figure 5) which illustrate the attentive regions that radiologists describe.

IU X-Ray [94] serves as one of the most widely adopted English datasets by existing RRG studies, which is introduced by Indiana University. To construct the dataset, the authors first collect a series of chest X-Ray images and radiology reports from two large hospital systems within the Indiana Network for Patient Care database. Then, they remove texts in all radiology reports that reflect the patients' privacy, e.g., patient names, hospital numbers, and complete dates, and filter out the radiographs with potential identifiers such as teeth, partial jaw, jewelry, partial skull, and specific medical devices. Finally, they manually match these processed radiographs with the corresponding reports, and construct them into radiograph-report pairs for public release, obtaining the final public version of IU X-Ray with 7,470 radiographs and 3,955 reports. Additionally, the authors use MeSH[5] to annotate medical terms in all radiology reports, and use RadLex[6] to cover all radiology terminology thatoutsidee of MeSH's purview, resulting in 101 MeSH labels and 76 RadLex labels for all impression and

4. We collect the examples of IU X-Ray, MIMIC-CXR, MIMIC-CXR-JPG from their public datasets, and select the examples of PEIR Gross from http://peir.path.uab.edu/library/. For CX-CHR and COV-CTR, we present the cases in https://www.researchgate.net/figure/Two-samples-from-CX-CHR-and-COV-CTR-datasets-Red-bounding-boxes-annotated-by-a_fig1_342027618. Since we have no access to the JLiverCT dataset, the example of JLiverCT is not presented.

5. https://www.nlm.nih.gov/mesh/meshhome.html
6. http://radlex.org/

findings sections of radiology reports, respectively. They also classify all reports into two categories, i.e., normal and abnormal, which contain 2,470 abnormal reports and 1,485 reports, respectively, so that related studies are able to locate these reports according to their demands. Normally, existing RRG studies follow the data splits proposed by Jing *et al.* [5], which partition IU X-Ray into the training, validation, and testing sets with the ratio of 7:1:2, which results in 5,229, 747, and 1,494 radiograph-report pairs, respectively.

MIMIC-CXR [95] is the largest English dataset with chest X-ray radiographs and radiology reports for existing RRG studies, where the dataset is also a commonly used testing benchmark in existing RRG studies. The dataset is collected from the Beth Israel Deaconess Medical Center (BIDMC) and is designed to support a wide body of research in medicine, including medical image understanding, natural language processing, and clinical decision support. To construct the dataset, the authors collect 473,075 radiographs and 377,110 reports of 65,379 patients from 2011 to 2016 in the emergency department of BIDMC, and match each radiology report with corresponding radiographs, resulting in 270,790 studies in the public version. Specifically, radiographs are stored in the format of digital imaging and communications in medicine (DICOM), which contains meta-data associated with one or more radiographs so as to facilitate related studies. These radiographs are de-identified by removing patient identifiers, with radiologically relevant information such as orientation retained. For radiology reports, the free-text reports are de-identified by replacing key information (e.g., personal health information (PHI) length) with underscores (i.e., "_ _ _"). Besides, they employ medical term tagger (e.g., CheXpert [68]) to annotate medical labels for all radiology reports in MIMIC-CXR, where these labels are classified into 14 categories that are associated with thoracic diseases and support devices. There are studies [96], [97] that construct datasets that are extended from MIMIC-CXR, which emphasize improving the data formats or focus on challenging cases of generating reports for chest X-ray radiographs. Particularly, **MIMIC-CXR-JPG** [96] is introduced from BIDMC based on MIMIC-CXR. The dataset collects 377,110 radiographs and 227,835 reports, and transforms all radiographs from DICOM to JPEG format, so as to facilitate related studies in processing these data. Moreover, they adopt a more powerful medical term tagger NegBio [119] to annotate medical labels for all radiology reports. In addition, **MIMIC-ABN** [97] dataset is a subset of MIMIC-CXR with emphasis on reports for abnormal radiographs. The proposed dataset selects all the abnormal radiographs and corresponding radiology reports from MIMIC-CXR, so as to train neural networks in describing complicated abnormalities of radiographs, and it contains 38,551 radiograph-report pairs in total.

3.2 Other Datasets

In addition to benchmark datasets, there are some other specific datasets used in existing RRG studies. These datasets are relatively small in scale and diversified compared to benchmark ones, and usually target on particular applications of RRG, e.g., generating reports for COVID patients' chest radiographs and liver CT images. Specifically, we mainly introduce three of them, **COV-CTR** [24] **JLiverCT** [20], and **PEIR Gross** [5] in this paper and illustrate the details of each one in the following texts.

COV-CTR [24] is a Chinese dataset that focuses on automatically generating radiology reports for CT images of COVID-19 patients. The dataset is built based on the COVID-CT dataset [121]. To construct COV-CTR, the authors invite three radiologists from the First Affiliated Hospital of Harbin Medical University to write professional reports for the radiographs of COVID-CT patients. Finally, the dataset consists of 728 radiographs (349 for COVID-19 and 379 for non-COVID) and their associated Chinese reports. Although the scale COV-CTR is relatively small compared to other datasets for RRG, the study provides enlightening heuristics for adapting RRG to particular diseases, especially the ones that meet real-time clinical demands.

JLiverCT [20] is a dataset that focuses on generating radiology reports for liver CT images. The authors collect radiographs and the corresponding Japanese radiology reports of liver lesions from a hospital, and employ professional radiologists to manually annotate the finding labels in the reports, resulting in 65 categories of finding labels. For each radiology report, sentences that do not describe the CT images' findings are omitted so as to avoid violating patients' privacy. Finally, JLiverCT dataset contains 1,083 liver CT images and the corresponding radiology reports. While most datasets analyze chest X-ray images, the JLiverCT dataset serves as a novel database that emphasizes CT images, which provides insights for the RRG community to address other possible data formats such as MRI imaging.

PEIR Gross [5] emphasizes producing medical reports for images of gross lesions (i.e., lesion regions visible to human naked eyes). **PEIR Gross** consists of medical images and their corresponding medical records from the Pathology Education Informational Resource (PEIR) digital library[7]. The dataset contains 7,442 radiographs of different gross lesions, e.g., abdomen, adrenal, chest, gastrointestinal, etc., which are collected from 21 PEIR pathology sub-categories, along with their associated reports. Since each report in PEIR Gross only contains one descriptive sentence for the corresponding radiograph, which significantly differs from normal reports that are in long contexts, the dataset does not attract much attention from recent studies for RRG.

4 Evaluation Metrics

Evaluation metrics are crucial for RRG for providing measurements on the quality of the produced radiology reports from different approaches and ensure a fair comparison among counterparts. Similar to other AI research, prevailing approaches adopt automatic metrics RRG evaluation through comparing the generated reports with the gold standard references (doctor-written reports). In general, metrics for this task are categorized into five types, including natural language generation (NLG), clinical efficacy (CE), standard image captioning (SIC), embedding-based, and task-specific features-based metrics, where NLG metrics and CE metrics are the most widely adopted ones in existing approaches. Table 3 presents an overview of all evaluation

7. http://peir.path.uab.edu/library/

TABLE 3
An overview of different evaluation metrics with respect to their metric categories, original task, used additional toolkit, and the input type.

Metric Category	Metric	Original Task	Additional Toolkit	Input Type
NLG	BLEU [80]	Translation	-	Report
	METEOR [81]	Translation	-	Report
	ROUGE [82]	Summarization	-	Report
CE	Precision	-	CheXpert	Labels
	Recall	-	CheXpert	Labels
	F1	-	CheXpert	Labels
SIC	CIDEr [84]	Captioning	-	Report
Embedding	BERTScore [120]	Text Similarity	BERT	Embeddings
TSF	Factually-oriented [30]	RRG	-	Entities
	MRIQI [16]	RRG	NegBio, CheXpert	Labels
	CO [20]	RRG	-	Labels
	nTKD [19]	RRG	-	Labels

metrics with respect to their metric categories, original task, used additional toolkit and the input type. Details of the aforementioned metrics are illustrated as follows.

4.1 Natural Language Generation Metrics

NLG metrics are initially designed for different natural language processing (NLP) tasks, which are adopted for RRG to measure the quality of the generated reports. Conventional NLG metrics consist of bilingual evaluation understudy (**BLEU**) [80], metric for evaluation of translation with explicit ordering (**METEOR**) [81], and recall-oriented understudy for gisting evaluation (**ROUGE**) [82].

BLEU [80] is introduced for machine translation, which measures the n-gram precision of generated tokens. Given the generated report with N_r tokens and the corresponding gold standard report with N_{r^*} tokens, the BLEU score of n-gram BLEU-n is computed by

$$\text{BLEU-}n = l_{BP} \cdot \exp\left(\sum_{i=1}^{\mathcal{N}} W_i \cdot \log P_i\right) \tag{11}$$

where $\mathcal{N}$ refers to the number of n-grams, W_i represents the weight of the i-th n-gram, and P_i denotes the precision of the i-th n-gram. Herein, l_{BP} represents the brevity penalty (BP) coefficient to punish the score when the generated report is too short, where $l_{BP} = \exp\left(1 - \frac{N_r}{N_{r^*}}\right)$ if $N_r \leq N_{r^*}$ and $l_{BP} = 1$ if $N_r > N_{r^*}$. BLEU-n with n increases from 1 up to 4 is usually used for the evaluation of RRG approaches, which measures the accuracy and coherence of the generated report to a certain extent.

METEOR [81] is designed for machine translation, which computes the recall of matching uni-grams from tokens in produced and gold standard reports according to their exact stemmed form and meaning. Given N^u as the number of uni-grams shared by the generated reports $\widehat{\mathcal{R}}$ and gold standard report $\mathcal{R}^*$, the uni-gram precision score P_u and uni-gram recall score R_u are computed by $P_u = \frac{N^u}{N^u_r}$ and $R_r = \frac{N^u}{N^u_{r^*}}$, where N^u_r and $N^u_{r^*}$ represent the numbers of uni-grams in $\widehat{\mathcal{R}}$ and $\mathcal{R}^*$, respectively. Then, the F-score is calculated using harmonic mean through $F = \frac{10 \cdot P \cdot R}{R + 9 \cdot P}$. Given the penalty coefficient p, the METEOR score is computed by

$$\text{METEOR} = F \cdot (1 - p) \tag{12}$$

where $p = 0.5 \cdot \frac{N_c}{N^u}$ and N_c denotes the number of chunks[8].

ROUGE-L [82] is proposed for summarization and applied for RRG with its variant, e.g., **ROUGE-L**, which measures the similarity between the generated and gold standard report based on their longest common subsequence (LCS) tokens. Given the generated report $\widehat{\mathcal{R}}$ with N_r tokens and gold standard report $\mathcal{R}^*$ with N_{r^*} tokens, their LCS tokens are first extracted as $R_{LCS}(\widehat{\mathcal{R}}, \mathcal{R}^*)$. Then, the precision score $P_{\text{ROUGE-L}}$ and the recall score $R_{\text{ROUGE-L}}$ are computed through $P = \frac{R_{LCS}(\widehat{\mathcal{R}}, \mathcal{R}^*)}{N_r}$ and $R = \frac{R_{LCS}(\widehat{\mathcal{R}}, \mathcal{R}^*)}{N_{r^*}}$, respectively. Finally, the ROUGE-L score is calculated through

$$\text{ROUGE-L} = \frac{(1 + \beta^2) \cdot R \cdot P}{R + \beta^2 \cdot P} \tag{13}$$

where β is a hyper-parameter that is usually set to a large number, and thus emphasizes the recall score [82].

4.2 Clinical Efficacy Metrics

CE metrics are widely utilized in existing approaches to evaluate their model performance in generating factually correct radiology reports. To compute CE metrics, medical labelers (i.e., NegBio [119], CheXpert [68], and MeSH) are firstly employed to label tokens in both generated report $\widehat{\mathcal{R}}$ and the gold standard one $\mathcal{R}^*$, where the obtained labels are related to thoracic diseases and support devices of 14 categories. Then, precision, recall, and F1 scores are computed by comparing the medical labels extracted from $\widehat{\mathcal{R}}$ and $\mathcal{R}^*$. Finally, the aforementioned scores are used as the results of the CE metrics, so as to measure the prediction accuracy of medical labels in the generated report.

8. A chunk is defined as a set of uni-grams that are adjacent in the generated report and the gold standard report.

4.3 Standard Image Captioning Metrics

Standard image captioning (SIC) metrics initially evaluate the quality of generated descriptions for natural images. Existing RRG approaches [5], [6], [14], [15], [16], [17] mainly adopt consensus-based image description evaluation (**CIDEr**) [84] following the standard image captioning evaluation settings upon the RRG task. **CIDEr** measures the quality of generated reports through cosine similarity. Specifically, CIDEr first calculates the n-gram cosine similarities between the produced reports and gold standard ones, obtaining the n-gram scores with n ranges from 1 to 4. Then, it computes the mean value over all n-gram similarities and uses the resulting value as the final CIDEr score.

4.4 Embedding-based Metrics

Embedding-based metrics measure the quality of the generated report through latent distance. In doing so, feature extractors (e.g., BERT [85]) are mainly adopted to project the generated and gold standard reports to the same latent space and compute the distance between their corresponding latent representations as the metric score. Specifically, current RRG approaches [20] utilize BERTScore [120] for the RRG task. They employ large-scale pre-trained language model (i.e., BERT [85]) to help the evaluation of similarity between generated and gold standard reports. Given the generated report $\widehat{\mathcal{R}}$ and gold standard one $\mathcal{R}^*$, BERTScore uses BERT to encode $\widehat{\mathcal{R}}$ and $\mathcal{R}^*$ into their corresponding latent representations $\mathbf{h}_{\widehat{\mathcal{R}}}$ and $\mathbf{h}_{\mathcal{R}^*}$. Then, it computes the cosine similarity matrix $\mathcal{M}$ between $\mathbf{h}_{\widehat{\mathcal{R}}}$ and $\mathbf{h}_{\mathcal{R}^*}$, and selects the maximum similarity score c_{max} among $\mathcal{M}$. Finally, BERTScore uses Inverse Document Frequency (IDF) to provide weights for each different token, and compute the precision, recall, and F1 score based on c_{max}, where the F1 score is utilized as the final score of BERTScore.

4.5 Task-Specific Features-based Metrics

Task-specific features-based metric aims to utilize the task-specific features (TSF) of RRG to provide comprehensive evaluation from dedicated aspects. These metrics consist of **factually-oriented metric** [30], medical image report quality index (**MIRQI**) [16], content ordering (**CO**) [20], and normalized key term distance (**nTKD**) [19], where the details of them are illustrated in the following texts.

Factually-oriented Metric [11] is introduced to measure the factual correctness and completeness of the generated radiology reports, which contains the exact entity match reward $\mathbf{fact_{ENT}}$ and the entailing entity match reward $\mathbf{fact_{ENTNLI}}$. Specifically, $\mathbf{fact_{ENT}}$ applies a named entity recognizer (NER) (i.e., Stanza [76]) to the generated report $\widehat{\mathcal{R}}$ and the gold standard report $\mathcal{R}^*$, resulting in two sets of entities which are denoted as $E_{\widehat{\mathcal{R}}}$ and $E_{\mathcal{R}^*}$, respectively. Then, $\mathbf{fact_{ENT}}$ is computed as the harmonic mean of precision and recall between $E_{\widehat{\mathcal{R}}}$ and $E_{\mathcal{R}^*}$. Furthermore, $\mathbf{fact_{ENTNLI}}$ serves as an extension of $\mathbf{fact_{ENT}}$ with natural language inference (NLI). Based on the standard process of $\mathbf{fact_{ENT}}$, $\mathbf{fact_{ENTNLI}}$ assumes that an entity e of $E_{\widehat{\mathcal{R}}}$ is not considered correct if it presents in $E_{\mathcal{R}^*}$. To consider e as a valid entity, the sentence $s_{\widehat{\mathcal{R}}}$ containing e must not present a contradiction with its counterpart sentence $s_{\mathcal{R}^*}$ in $\mathcal{R}^*$, where $s_{\mathcal{R}^*}$ refers to the sentence with the highest BERTScore [120] in $\mathcal{R}^*$. Herein, $\mathbf{fact_{ENTNLI}}$ utilizes a pre-trained NLI model [30] to judge whether $s_{\widehat{\mathcal{R}}}$ is a contradiction of $s_{\mathcal{R}^*}$.

MIRQI [16] is proposed to evaluate the quality of reports by measuring the degree of match between the generated reports and the gold standard from the perspectives of the disease keywords and their associated attributes. To calculate the metric, medical labeling toolkits are adopted (e.g., NegBio [119] and CheXpert [68]) to extract two sets of disease words $W_{\widehat{\mathcal{R}}}$ and $W_{\mathcal{R}^*}$ from both generated reports $\widehat{\mathcal{R}}$ and gold standard reports $\mathcal{R}^*$, respectively. The extracted disease words $W_{\widehat{\mathcal{R}}}$ and $W_{\mathcal{R}^*}$ compose graphs $G_{\widehat{\mathcal{R}}}$ and $G_{\mathcal{R}^*}$ for $\widehat{\mathcal{R}}$ and $\mathcal{R}^*$, where the child nodes of $G_{\widehat{\mathcal{R}}}$ and $G_{\mathcal{R}^*}$ are obtained as the attributes that represent the features of diseases. Given $G_{\widehat{\mathcal{R}}}$ and $G_{\mathcal{R}^*}$, the precision, recall, and F1 score are computed and used for evaluation by comparing $W_{\widehat{\mathcal{R}}}$ and $W_{\mathcal{R}^*}$ and their associated attributes.

CO [20] is designed to quantify the consistency of the description order of the generated reports. It indicates a chronological order of the report contents, and reveals their clinical importance regarding that important findings of a report are likely to be written in the earlier parts. Given the generated report $\widehat{\mathcal{R}}$ and the gold standard report $\mathcal{R}^*$, the metric first adopts medical labelers (e.g. CheXpert [68]) to extract finding labels from $\widehat{\mathcal{R}}$ and $\mathcal{R}^*$, which are denoted as l and l^*, respectively. Then, the CO score is computed by the Damerau-Levenshtein distance between l and l^*.

nTKD [19] aims to judge whether sentences in the generated report contain all observed diseases and their detailed descriptive information. To compute nTKD, the metric firstly employs medical labelers to extract medical labels from the generated report $\widehat{\mathcal{R}}$ and gold standard report $\mathcal{R}^*$, resulting in two sets of labels l and l^*, respectively. Then, the nKTD score s_{nKTD} is computed through Hamming distance $d_{hamming}(\cdot)$ between l and l^*, formulated by $s_{nKTD} = \frac{d_{hamming}(l,l^*)}{N_l}$ with N_l as the number of all labels.

5 Results and Analysis

In this section, we summarize the experiment settings of existing RRG studies and conduct a performance comparison on some representative approaches, with analyses of the impact of different approaches according to their model architecture and method categories.

5.1 Experiment Settings

To assess the effectiveness of RRG, current approaches primarily utilize a range of experiments focusing on various aspects, including quantitative and qualitative evaluations, alongside other specific methods to evaluate its performance. Quantitative evaluation mainly adopts mainstream metrics (i.e., NLG and CE metrics) for evaluation, which indicates the quality of generated reports with computed scores of different metrics. Qualitative evaluation attempts to present visualization or analyses based on the produced results of RRG model. For example, most existing RRG studies conduct case studies to present the qualitative evaluation, where they select specific radiographs and their corresponding reports from RRG datasets for

reference, and directly demonstrate the outputs of different RRG approaches, e.g., radiology reports and medical terms. Moreover, some attention-based approaches [5], [8], [12], [18], [19], [21] visualize the internal attention maps, so as to demonstrate the image-text mappings between specific regions of radiographs and medical terms in the generated reports established by their models. Particularly, some studies perform method-specific evaluations based on their method categories. Particularly, knowledge enhancement-based approaches [10], [15], [16], [17], [20], [22], [25], [26], [43] evaluate the accuracy of the predicted graph nodes or medical terms for a comprehensive evaluation of their approaches. RRG studies [23], [28] that employ regional visual features manage to present the visualization of predicted bounding boxes, so as to present the performance of the fine-tuned object detectors of their approaches for RRG.

5.2 Performance Comparison

To analyze the impact of different model architectures and method categories, we present a series of performance comparisons upon existing RRG studies. Table 4 presents the performance comparison on IU X-Ray [94] among main approaches with different model architectures w.r.t NLG metrics. Table 5, 6, and 7 report the performance comparison of main approaches upon MIMIC-CXR [95], IU X-Ray [94], and other datasets (i.e., CX-CHR [5], COV-CTR [24], MIMIC-ABN [97] based on the method categories of different studies, respectively. In the following texts, we first analyze the impacts of different model architectures and summarize the influences of different method categories according to their compared performance on different datasets.

5.2.1 Comparison of Model Architecture

The model architectures of existing RRG studies are categorized into three main groups, including CNN-LSTM architecture, GCN-LSTM architecture, GCN-Transformer architecture, and Transformer-Transformer architecture, whose details are illustrated as follows. Specifically, CNN-LSTM architecture is adopted by many approaches [5], [19], [28]. It uses CNN or its variants to encode radiographs into latent features and utilizes LSTM to decode sentences in the final report. These approaches provide heuristics for RRG, and contain significant room for performance improvement. Some studies [15], [16], [17], [25], [26], [43] manage to introduce knowledge graph to facilitate the report generation process, which motivates the requirements of processing graph-based information for RRG. Therefore, GCN-based encoders are proposed and lead to improvements on CNN-LSTM-based approaches, demonstrating the effectiveness of knowledge graphs for RRG, where such encoders help the model build relationships between diseases and organs through encoding the corresponding nodes and edges of the knowledge graph for the LSTM decoder, thereby performing a more effective encoding process compared to CNN-based encoders. With the rising of Transformer-based architecture, recent studies [9], [10], [12], [13], [17], [18], [20], [21], [22], [23], [24], [26], [27], [43] employ Transformer for RRG, which is first utilized to improve the text decoder, and then employed to enhance both encoders and decoders. Specifically, GCN-Transformer architecture brings further improvements upon GCN-LSTM architecture, with Transformer-based decoders to handle the report generation process more effectively. Later, Transformer is also conducted for encoders and obtains the best performance over existing model architectures, where the multi-head attention mechanism in Transformer demonstrates its superiority for both encoding radiographs and generating radiology reports. This observation demonstrates that the multi-head attention mechanism currently serves as the best solution to model the long report generation process, where the self-attention layers learn the internal correlation of regions in radiographs as well as words in radiology reports.

TABLE 4
Comparison of existing approaches on IU X-Ray based on their encoder-decoder architecture, where the background of CNN-LSTM, GCN-LSTM, GCN-Transformer, and Transformer-Transformer architectures are highlighted in grey, red, green, and blue, respectively. Herein, the **best** result is highlighted in boldface.

Study	NLG Metrics					
	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L
Vinyals *et al.* [93]	0.216	0.124	0.087	0.066	-	0.306
Rennie *et al.* [32]	0.224	0.129	0.089	0.068	-	0.308
Lu *et al.* [33]	0.220	0.127	0.089	0.068	-	0.308
Jing *et al.* [14]	0.464	0.301	0.210	0.154	-	0.362
Jing *et al.* [5]	0.455	0.288	0.205	0.154	-	0.369
Li *et al.* [6]	0.438	0.298	0.208	0.151	-	0.322
Xue *et al.* [37]	0.464	0.358	0.270	0.195	0.274	0.366
Harzig *et al.* [39]	0.373	0.246	0.175	0.126	0.163	0.315
Wang *et al.* [28]	0.487	0.346	0.270	0.208	-	0.359
Liu *et al.* [29]	0.492	0.314	0.222	0.169	0.193	0.381
Ni *et al.* [19]	0.536	0.391	0.314	0.252	0.228	**0.448**
Yang *et al.* [55]	0.478	0.344	0.248	0.180	-	0.398
Avg. score	0.397	0.265	0.194	0.146	0.215	0.349
Zhang *et al.* [16]	0.441	0.291	0.203	0.147	-	0.367
Li *et al.* [15]	0.455	0.288	0.205	0.154	-	0.369
You *et al.* [10]	0.479	0.319	0.222	0.174	0.193	0.377
Avg. score	0.467	0.304	0.214	0.164	0.193	0.373
Chen *et al.* [8]	0.470	0.304	0.219	0.165	-	0.371
Nooralahzadeh *et al.* [9]	0.486	0.317	0.232	0.173	0.192	0.390
Liu *et al.* [17]	0.483	0.315	0.224	0.168	-	0.376
Li *et al.* [43]	-	-	-	0.163	0.193	0.383
Hou *et al.* [26]	0.510	0.346	0.255	0.195	**0.205**	0.399
Liu *et al.* [29]	0.492	0.314	0.222	0.169	0.193	0.381
Liu *et al.* [78]	0.473	0.305	0.217	0.162	0.186	0.378
Wang *et al.* [28]	0.487	0.346	0.270	0.208	-	0.359
Nooralahzadeh *et al.* [9]	0.486	0.317	0.232	0.173	0.192	0.390
Chen *et al.* [18]	0.475	0.309	0.222	0.170	0.191	0.375
You *et al.* [46]	0.484	0.313	0.225	0.173	0.204	0.379
Hou *et al.* [45]	0.232	-	-	-	0.101	0.240
Delbrouck *et al.* [11]	-	-	-	0.139	-	0.327
Wang *et al.* [47]	0.525	0.357	0.262	0.199	0.220	0.411
Qin *et al.* [21]	0.494	0.321	0.235	0.181	0.201	0.384
Yu *et al.* [48]	0.457	0.305	0.216	0.171	-	0.391
Nicolson *et al.* [56]	0.473	0.304	0.224	0.175	0.200	0.376
Yan *et al.* [49]	0.482	0.313	0.232	0.181	0.203	0.381
Wang *et al.* [50]	0.505	0.340	0.247	0.188	0.208	0.382
Kong *et al.* [52]	0.484	0.333	0.238	0.175	0.207	0.415
Tanwani *et al.* [53]	**0.580**	**0.440**	**0.320**	**0.270**	-	-
Wang *et al.* [54]	0.496	0.319	0.241	0.175	-	0.377
Wang *et al.* [36]	0.505	0.345	0.243	0.176	205	0.396
Huang *et al.* [13]	0.525	0.360	0.251	0.185	**0.242**	0.409
Wang *et al.* [12]	0.483	0.322	0.228	0.172	0.192	0.380
Avg. score	0.480	0.329	0.223	0.200	0.196	0.414

5.2.2 Comparison of Method Categories

To analyze the impact of method categories in existing studies, we first illustrate the impacts of each specific category, and then compare them across different categories to discuss the effect brought by different modality enhancement. For reference, we report the performance of image captioning approaches on RRG datasets. In Figure 6, we demonstrate some examples of generated reports by visual-only, textual-only, and cross-modal approaches with respect to different characteristics. Details are introduced in the following texts.

TABLE 5
Results of representative RRG models on MIMIC-CXR with respect to NLG and CE metrics, where image captioning approaches, visual-only approaches, textual-only approaches, and cross-modal approaches are marked by gray, red, green, and blue backgrounds, respectively.

	NLG Metrics						CE Metrics		
Study	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	Precision	Recall	F1
Vinyals *et al.* [93]	0.299	0.184	0.121	0.084	0.124	0.263	0.249	0.203	0.204
Rennie *et al.* [32]	0.325	0.203	0.136	0.096	0.134	0.276	0.322	0.239	0.249
Lu *et al.* [33]	0.299	0.185	0.124	0.088	0.118	0.266	0.268	0.186	0.181
Anderson *et al.* [34]	0.317	0.195	0.130	0.092	0.128	0.267	0.320	0.231	0.238
Avg. score	0.310	0.192	0.128	0.090	0.126	0.268	0.290	0.215	0.218
Tanida *et al.* [23]	0.373	0.249	0.175	0.126	0.168	0.264	0.461	0.475	0.447
Wang *et al.* [36]	0.363	0.235	0.164	0.118	0.136	0.301	-	-	-
Avg. score	0.368	0.242	0.170	0.122	0.152	0.283	-	-	-
Chen *et al.* [8]	0.353	0.218	0.145	0.103	0.142	0.277	0.333	0.273	0.276
Nooralahzadeh *et al.* [9]	0.378	0.232	0.154	0.107	0.145	0.272	0.240	0.428	0.308
Liu *et al.* [17]	0.360	0.224	0.149	0.106	0.149	0.284	-	-	-
Nishino *et al.* [20]	-	-	-	0.122	-	0.168	-	-	-
Dalla Serra *et al.* [41]	0.363	0.245	0.178	**0.136**	0.161	**0.313**	0.428	0.459	0.443
Li *et al.* [43]	-	-	-	0.109	0.150	0.284	-	-	-
Hou *et al.* [26]	0.407	0.256	0.172	0.123	0.162	0.293	0.416	0.418	0.385
Avg. score	0.372	0.235	0.160	0.115	0.151	0.270	0.329	0.395	0.353
Nishino *et al.* [7]	0.363	0.231	0.163	0.121	-	0.267	-	-	-
Chen *et al.* [18]	0.353	0.218	0.148	0.106	0.142	0.278	0.334	0.275	0.278
You *et al.* [46]	0.378	0.235	0.156	0.112	0.158	0.283	-	-	-
Yan *et al.* [27]	0.373	-	-	0.107	0.144	0.274	0.385	0.274	0.294
Ni *et al.* [19]	0.372	0.241	0.168	0.123	**0.190**	0.335	-	-	-
Delbrouck *et al.* [11]	-	-	-	0.116	-	0.259	-	-	-
Qin *et al.* [21]	0.381	0.232	0.155	0.109	0.151	0.287	0.342	0.294	0.292
Wang *et al.* [47]	0.344	0.215	0.146	0.105	0.138	0.279	-	-	-
Yu *et al.* [48]	0.347	0.235	0.149	0.106	-	0.280	0.447	0.593	0.503
Nicolson *et al.* [56]	0.393	0.248	0.171	0.127	0.155	0.286	0.367	0.418	0.391
Yan *et al.* [49]	0.356	0.222	0.151	0.111	0.140	0.280	0.353	0.310	0.297
Wang *et al.* [50]	0.395	0.253	0.170	0.121	0.147	0.284	-	-	-
Wang *et al.* [51]	0.413	**0.266**	**0.186**	**0.136**	0.168	0.286	**0.482**	**0.563**	**0.519**
Kong *et al.* [52]	**0.423**	0.261	0.171	0.116	0.170	0.298	-	-	-
Wang *et al.* [54]	0.351	0.223	0.157	0.118	-	0.287	-	-	-
Yang *et al.* [55]	0.362	0.251	0.188	0.143	-	0.326	-	-	-
Huang *et al.* [13]	0.393	0.243	0.159	0.113	0.160	0.285	0.371	0.318	0.321
Wang *et al.* [12]	0.386	0.250	0.169	0.124	0.152	0.291	0.364	0.309	0.311
Avg. score	0.378	0.239	0.161	0.116	0.155	0.285	0.383	0.373	0.356

Visual-only Approaches mostly adopt global visual features for RRG, where several studies employ regional visual features or global-regional aggregated ones for further enhancement. By comparing whether using regional visual features for RRG, Tanida *et al.* [23] that employ regional visual features obtain improvements over other approaches using global visual features (e.g., Wang *et al.* [12]), suggesting the effectiveness of utilizing regional visual features to provide fine-grained visual information for the report generation process. Notably, this approach obtains significantly improved performance on CE metrics over approaches that use global visual features, which indicates that regional visual features are able to facilitate the overall pipeline in matching particular regions in radiographs to medical terms in radiology reports, as is demonstrated in Figure 6. Also, global-regional aggregated features also demonstrate its effectiveness in improving RRG, where Li *et al.* [24] reveal better performance over Li *et al.* [15] on CX-CHR. This observation verifies the validity of considering both global and regional views of radiographs for RRG, which is similar to the report-writing process of physicians. Nevertheless, the number of visual-only approaches is still limited, while existing studies mainly focus on employing textual information and enhanced radiograph-report alignment to assist RRG, so the paradigm of visual-only approaches still has large potentials for further enhancement.

Textual-only Approaches mainly enhance RRG through knowledge enhancement, report structuralization, and progressive report generation, with several observations as follows. Knowledge enhancement-based approaches [10], [15], [16], [17], [20], [22], [25], [26], [43] play a dominant role in existing RRG studies. They employ textual information (e.g., medical terms, entities, and relations, as well as knowledge graph) to facilitate RRG. Particularly, some studies that adopt medical terms (e.g., You *et al.* [10]) demonstrate significant improvement compared to vanilla image captioning approaches, indicating that the report generation process is effectively guided through disease classification and medical terms prediction. Knowledge graph-based approaches (e.g., Zhang *et al.* [16], Li *et al.* [43], and Hou *et al.* [26]) are one of the most mainstream approaches since the employment of deep learning in RRG, where

TABLE 6
Comparisons of existing RRG approaches on the test sets of IU X-Ray with respect to NLG metrics, where the image captioning approaches, textual-only approaches, and cross-modal approaches are highlighted in gray, green, and blue, backgrounds, respectively.

Study	NLG Metrics					
	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L
Vinyals *et al.* [93]	0.216	0.124	0.087	0.066	-	0.306
Rennie *et al.* [32]	0.224	0.129	0.089	0.068	-	0.308
Lu *et al.* [33]	0.220	0.127	0.089	0.068	-	0.308
Avg. score	0.220	0.127	0.088	0.067	-	0.307
Wang *et al.* [36]	0.505	0.345	0.243	0.176	205	0.396
Li *et al.* [15]	0.455	0.288	0.205	0.154	-	0.369
Jing *et al.* [14]	0.464	0.301	0.210	0.154	-	0.362
Harzig *et al.* [39]	0.373	0.246	0.175	0.126	0.163	0.315
Zhang *et al.* [16]	0.441	0.291	0.203	0.147	-	0.367
Chen *et al.* [8]	0.470	0.304	0.219	0.165	-	0.371
Nooralahzadeh *et al.* [9]	0.486	0.317	0.232	0.173	0.192	0.390
Liu *et al.* [17]	0.483	0.315	0.224	0.168	-	0.376
You *et al.* [10]	0.479	0.319	0.222	0.174	0.193	0.377
Li *et al.* [43]	-	-	-	0.163	0.193	0.383
Hou *et al.* [26]	0.510	0.346	0.255	0.195	0.205	0.399
Avg. score	0.474	0.310	0.221	0.166	0.196	0.377
Jing *et al.* [5]	0.455	0.288	0.205	0.154	-	0.369
Li *et al.* [6]	0.438	0.298	0.208	0.151	-	0.322
Liu *et al.* [29]	0.492	0.314	0.222	0.169	0.193	0.381
Liu *et al.* [78]	0.473	0.305	0.217	0.162	0.186	0.378
Wang *et al.* [28]	0.487	0.346	0.270	0.208	-	0.359
Chen *et al.* [18]	0.475	0.309	0.222	0.170	0.191	0.375
You *et al.* [46]	0.484	0.313	0.225	0.173	0.204	0.379
Hou *et al.* [45]	0.232	-	-	-	0.101	0.240
Ni *et al.* [19]	0.536	0.391	0.314	0.252	0.228	**0.448**
Delbrouck *et al.* [11]	-	-	-	0.139	-	0.327
Qin *et al.* [21]	0.494	0.321	0.235	0.181	0.201	0.384
Wang *et al.* [47]	0.525	0.357	0.262	0.199	0.220	0.411
Yu *et al.* [48]	0.457	0.305	0.216	0.171	-	0.391
Nicolson *et al.* [56]	0.473	0.304	0.224	0.175	0.200	0.376
Yan *et al.* [49]	0.482	0.313	0.232	0.181	0.203	0.381
Kong *et al.* [52]	0.484	0.333	0.238	0.175	0.207	0.415
Wang *et al.* [50]	0.505	0.340	0.247	0.188	0.208	0.382
Tanwani *et al.* [53]	**0.580**	**0.440**	**0.320**	**0.270**	-	-
Wang *et al.* [54]	0.496	0.319	0.241	0.175	-	0.377
Yang *et al.* [55]	0.478	0.344	0.248	0.180	-	0.398
Huang *et al.* [13]	0.525	0.360	0.251	0.185	**0.242**	0.409
Wang *et al.* [12]	0.483	0.322	0.228	0.172	0.192	0.380
Avg. score	0.478	0.331	0.241	0.183	0.198	0.374

these approaches also obtain competitive performance with medical term-based ones. Besides, approaches [15], [41] that adopt structuralized information (e.g., report templates) to facilitate RRG demonstrate competitive performance as well as knowledge-enhanced ones. Notably, the improvements of these RRG models over general image captioning approaches are particularly on BLEU-3 and BLEU-4 metrics, which indicates that the document-level guidance is able to pilot the text decoder in generating more coherent reports. Moreover, the progressive report generation-based approaches (e.g., Nooralahzadeh *et al.* [9]) also outperform image captioning approaches, with the two-stage generation paradigm indicating the effectiveness of generating reports in long contexts. Nevertheless, the performance gain of progressive report generation is less significant than other categories, suggesting that the impact of different text guidance upon the generation process varies according to the granularity level of guidance, where word-level information (e.g., medical terms and knowledge graphs) provides fine-grained guidance in generating descriptive texts, and document-level information (e.g., report template, related report, and intermediate result) puts emphasis on the construction of the entire radiology report.

Cross-modal Approaches generally all obtain competitive performance owing to their privileging model designs, including reinforcement learning, representation weighting, and architecture enhancement ones, respectively. In detail, reinforcement learning-based approaches [7], [11], [14], [21], [28], [78] that optimize the training objectives improve the RRG performance through carefully designed strategies such as reinforcement learning rewards and additional supervision signals, which obtain significant improvements over the image captioning approaches with standard cross-entropy loss. This observation illustrates that there are still improvements for RRG to design dedicated objectives, so that the cross-modal alignment radiographs and reports are enhanced to facilitate RRG. Representation weighting-based approaches [5], [8], [12], [13], [17], [18], [19], [21], [29] also outperform previous image captioning approaches, where the attention mechanisms [5], [19], [29] learn to weight the representation in different modalities, and the memory networks are optimized to establish the image-text matching between radiographs and reports. With particular model architecture designed for RRG, the designed RRG models (e.g., [13] and [12]) show promising performance, where RRG models are more capable of modeling the long text generation process than image captioning ones. The aforementioned observations illustrate the effectiveness of improving the cross-modal alignment based on whether additional modifications upon model architecture and optimization strategies are needed, which is able to provide references to adapt existing RRG studies to other applications in the future based on their application-specific characteristics.

TABLE 7
Comparisons of existing RRG approaches on the test sets of CX-CHR, COV-CTR, and JLiverCT datasets, with respect to NLG metrics.

Study	Dataset	NLG Metrics				
		BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L
Li *et al.* [15]	CX-CHR	0.673	0.588	0.532	0.473	0.618
Li *et al.* [24]	CX-CHR	**0.686**	**0.608**	**0.558**	**0.523**	**0.641**
Wang *et al.* [28]	COV-CTR	**0.810**	**0.766**	**0.721**	**0.679**	**0.790**
Li *et al.* [24]	COV-CTR	0.712	0.659	0.611	0.570	0.746
Yan *et al.* [27]	MIMIC-ABN	0.256	-	-	0.067	0.241
Nishino *et al.* [20]	JLiverCT	-	-	-	0.466	0.592

Comparison across Different Approach Categories reports the average performance of visual-only, textual-only, and cross-modal approaches on two main benchmark datasets (i.e., IU X-Ray [94] and MIMIC-CXR [95]), where we observe that the cross-modal approaches obtain the best performance on average, and the average score of the textual-only approach shows the second best, while the visual-only approach obtains the least performance. There are several conclusions drawn with respect to different categories of RRG studies. First, cross-modal approaches [5], [7], [8], [11], [12], [13], [14], [17], [18], [19], [21], [28], [29], [78] reveal promising performance for RRG, indicating that the pivotal roles of both radiographs and reports for RRG, especially the image-text mapping between them, where Figure 6 presents the visualization of the cross-modal mapping between specific regions and medical terms in the generated report. Second, textual-only approaches [10], [15], [16], [17], [20], [22], [25], [26], [27], [43] assist the report generation process with extra medical information (e.g., medical terms and related reports), which provide different granular levels of guidance for RRG, so that descriptive texts in the generated report

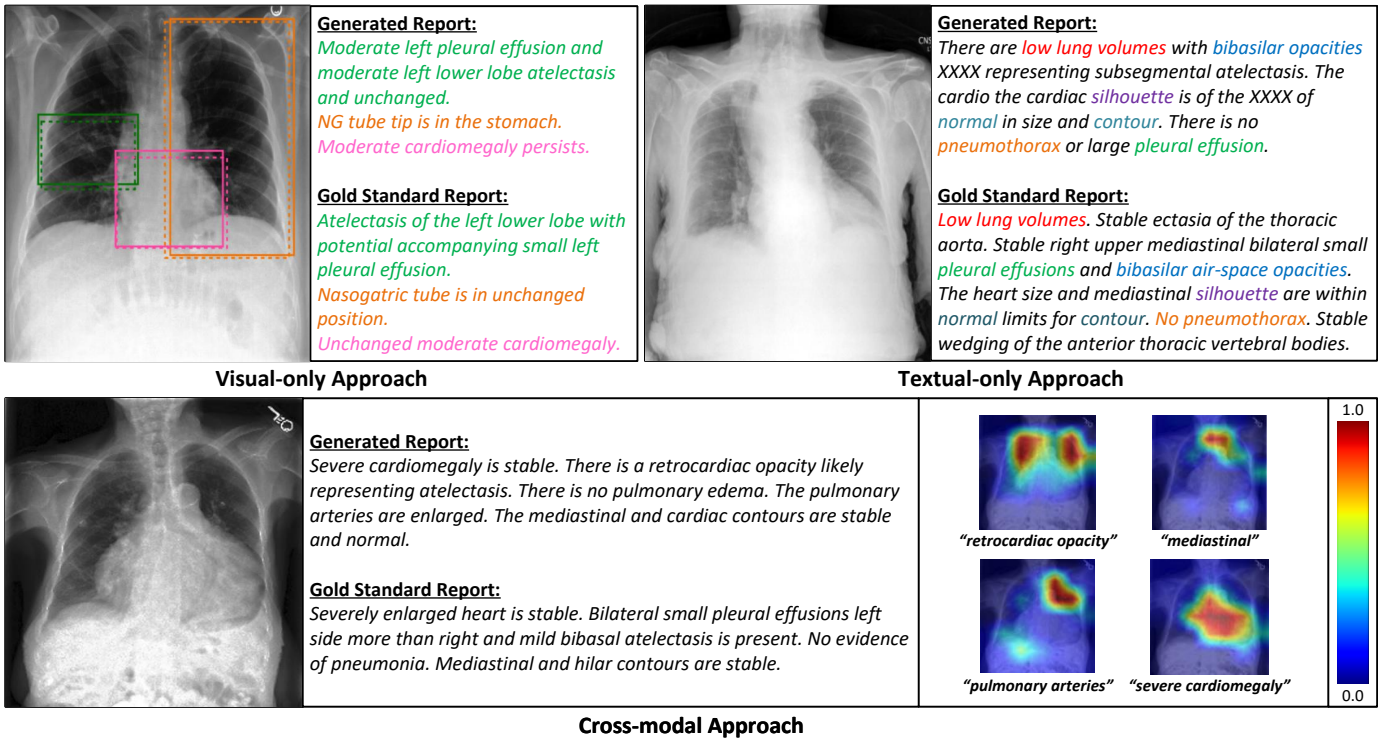


Fig. 6. Illustrations of the reports generated by visual-only, textual-only, and cross-modal approaches according to their input radiographs, where the gold standard reports are presented for reference. Herein, results of visual-only, textual-only, and cross-modal approaches are produced by Tanida *et al.* [23], You *et al.* [10], and Chen *et al.* [18], respectively. For visual-only approach, detected regions and the corresponding descriptive sentences are highlighted in the same color. For textual-only approach, medical terms shared by the generated report and the gold standard report are highlighted in the same colors. For cross-modal approach, we demonstrate the visualization of image-text mappings between particular regions of the radiograph and words/phrases from its generated report, with the color spectrum indicating the value of weight ranging from $[0, 1]$.

are well-aligned with the corresponding terms in the gold standard report as is shown in Figure 6. Compared to cross-modal approaches, textual-only approaches mainly facilitate RRG from the perspective of the textual modality, and fail to address the enhancement for processing radiographs, thereby obtaining fewer improvements compared to cross-modal approaches. Third, visual-only approaches [23], [24] obtain the least performance among all method categories for RRG, which indicates that extracting elaborated features from radiographs is a challenging task that is hard to effectively address, where more related studies in the future are expected to facilitate RRG from such perspective.

6 Challenges and Future Directions

As an interwoven research direction, RRG is intrinsically challenging that integrates the difficulties from both areas of AI and clinical medicine, which requires studies to consider the characteristics from AI models, algorithms as well as task-specific features in the medical domain, e.g., specific neural model architecture, different types of medical images, and the contextual specialties in reports, etc. Although the methodology review and experimental analysis of existing studies have demonstrated significant prosperity of RRG research over the past few years, many open challenges and opportunities still stand, and being worthy of, for putting sustainable attention upon, where data resources, model design, and evaluation protocols are three major directions, in our opinion, to be thoroughly investigated and developed to meet the requirements of real-world applications among many other important directions. We elaborate the details of our analysis on them in the following subsections.

6.1 Data Resources

Data resources, which directly influence the performance of the RRG models by their participating in the training and evaluation processes, play an essential role in RRG research. As mentioned in previous sections, existing approaches mainly conduct their experiments on benchmark datasets, which have the potential to be improved in scale, where their data instances are uniformed in types and lack diversity, making it difficult for the trained models to meet various real-world clinical needs. Thus, collecting and utilizing a wider range of data for this task is an anticipated direction for future research, which is elaborated as follows.

First, existing publicly available RRG datasets have the potential to be further improved in data quantity, which allows RRG approaches to achieve performance improvements through additional data. Although there are massive amounts of unlabeled data in radiograph datasets (e.g., NIH chest X-ray dataset [99]) and medical corpora (e.g., PubMed[9]), the supervised learninng paradigm of existing models are not able to utilize them for RRG owing to the fact of no radiograph-report pairs. Meanwhile, collecting datasets for RRG studies is normally resource-consuming and requires extensive data cleaning efforts to ensure data quality, which still cannot be fully automated and thus

9. https://pubmed.ncbi.nlm.nih.gov/download/

demands a considerable amount of human efforts. Therefore, to address the data quantity problem, approaches that utilize data with more efficiency are expected to improve RRG with data augmentation and unsupervised learning.

Second, existing datasets only support one type image-text mapping, which mainly consists of X-ray images and their corresponding reports. As a result, X-ray images provide limited information to physicians in clinical practice, and struggle to present comprehensive knowledge from additional aspects, e.g., the health status of particular organs, temporal information, and other views of patients. To address such limitations, other types of medical imaging are considered to provide enriched information for report generation. For example, magnetic resonance imaging (MRI) is able to present specific states of organs, ultrasound imaging demonstrates the temporal development of a patient's health condition, and 3D ultrasound imaging offers a holistic view of the patient to radiologists. Also, from the language perspective, existing radiology reports in public RRG datasets, especially the benchmark ones, are mainly written in English, whereas reports in other languages are also expected to promote the development of RRG research in different districts. Consequently, approaches that are capable of processing diverse radiographs and generating reports in multiple languages are expected to be studied.

Third, widely used datasets mainly focus on chest radiographs and reports, where there is still high demand for radiographs and reports for other body regions, particularly for pivotal parts (e.g., brains, bone joints), thereby resulting in difficulties in applying prevailing RRG approaches to clinical practice. Particularly in these regions, the imaging form and medical terminology in the corresponding reports have significant differences compared to the ones of chest X-ray images. Future research on more regions is thus highly expected to expand the application range of existing studies.

6.2 Model Design

As reviewed in previous sections, current deep learning-based RRG approaches are mainly restricted in using few well-established frameworks (e.g., CNN for visual encoding and LSTM/Transfomer for text decoding), which are, to some extent, highly integrated and not easy to be replaced and completely overturned, so that lead to the challenging fact that many studies are limited to incremental innovations in their methodologies. However, with the fast growing of AI, there are new possibilities in applying new techniques and even new paradigm to RRG, therefore many expectations are drawn in terms of model design for RRG.

Despite of using new models, the first one needs specific attention is the interpretability of RRG approaches. Although promising performance has been obtained on multiple datasets, existing studies still lack of reasonable explanation for how their model is designed and why they work. In trying to interpret the internal mechanism of their approaches, most studies [5], [8], [12], [18], [19], [21] analyze the cross-modal correlation between input radiographs and generated reports through heat map visualizations. However, such solutions are still unable or ambiguous to explicitly demonstrate the underlying mechanisms and reasons behind that contribute to RRG e.g., clinical theories, medical knowledge, diagnostic results, etc., where more explainable studies are expected in future RRG studies, especially in designing a model that is not performing in a black box.

From the aspect of radiograph processing, although existing studies employ different features for RRG, they struggle to represent radiographs in a detailed manner, leading to inaccuracies in generated reports. Even through some approaches adopt regional visual features [23], [28] to represent radiographs, the regions obtained by these approaches are still relatively coarse with a high degree of overlapping, leading to less accurate representations of radiographs and limited visual information for the report generation process. To address such limitations, scaled-up image segmentation models (e.g., segment anything (SAM) [122] and its variants for medical images [105], [106], [107], [108], [109], [110], [111], [112]), have shown potentials in extracting more fine-grained visual representations from radiographs, so that such models or similar designs should be able to provide more detailed visual information to enhance RRG.

For the generation process, the overwhelm majority of existing approaches use autoregressive models (i.e., LSTM/Transformer), and train them on public benchmark datasets from scratch. Although such approaches demonstrate outstanding performance, they are susceptible to the error propagation problem when incorrect texts are halfway produced during the report generation process. Therefore, non-autoregressive (non-AR) models (e.g., diffusion model [123]) are in the consideration for this task to offer an alternative paradigm through generating all words in a parallel manner, which has already demonstrated its effectiveness in other text generation tasks, e.g., sequence-to-sequence text generation [124] and image captioning [125], [126]. As a most important notice, by witnessing current circumstance on the development of large language models (LLMs) [127], [128], one should admit that LLMs have demonstrated their outstanding generation and in-context learning ability in few shot settings, providing a promising potential solution to generate elaborated radiology reports. Also, the in-context learning of LLM suggests possible variants of the RRG task, e.g., medical visual question answering and interactive report generation. Furthermore, recent vision-language models (VLMs) [129], [130], [131], [132] achieve impressive understanding and generation abilities through aligning vision backbone models (e.g., ViT) and LLMs, therefore showing their strength to enhance the radiograph-report alignment for RRG. For example, recent studies such as XrayGPT [133] adopt LLMs for radiology report summarization that generates a summary according to the input radiograph and the findings section of the report, which already proves the trend of introducing LLM for RRG. However, one possible concern for applying LLM to RRG is the domain adaptation problem, which requires adapting the knowledge of LLM from the general domain to the particular medical domain, especially with a limited amount of medical data. A Challenging mission worth dedicated research is thus how to effectively perform such adaptation, since domain mismatch is a difficult gap to overcome in applying LLMs to different real-world scenarios even though they possess strong few-shot learning ability.

6.3 Evaluation Protocols

Existing metrics for RRG evaluation generally follow the conventional settings of the AI field with using several routine measurements, in either using a reference report or dedicated labels to compare with the results from different models on the textual side. One significant limitation of doing so is the lack of cross-modal consideration with specific assessment on how good that particular regions in a radiograph and the texts in the report are associated with each other. Although attention heat map is widely adopted to present the alignment between specific regions and medical labels in attention- and Transformer-based approaches [5], [8], [12], [18], [19], [21], there still requires quantitative evaluation to demonstrate such cross-modal connections. In other multi-modal tasks, e.g., image captioning, there are already some automatic metrics to assess the cross-modal correlation between generated texts and the input images. For example, CLIPScore [134] is used to evaluate the alignment between texts and images considering the distance of their features in the same semantic space, which offers a reference to measure the cross-modal relations for RRG. From another aspect, existing metrics require references to measure the quality of the generated reports, yet it is generally hard to have such references standby in real-world applications. Therefore, activity involving human evaluation has provided a practical option to measure the quality of radiology reports as well as cross-modal alignment between particular regions in radiographs and the texts in reports, and demonstrated significant potential in a series of scenarios such as telemedicine consultation. For example, A/B testing for physicians' references on generated reports is a promising solution that presents radiologists with comparisons by different RRG approaches as well as providing feedbacks the reports' quality and clinical validity according to human preferences.

7 Conclusion

RRG plays a pivotal role in applying AI techniques to the clinical medicine field. In panoramically investigating the development of recent RRG research, in this paper, we propose to categorize existing RRG studies into three groups based on their enhancing modalities, namely visual-only, textual-only, and cross-modal approaches. Then, we summarize the evaluation standards for RRG, consisting of different datasets and evaluation metrics. Afterwards, we introduce the experiment settings of different approaches and their results, as well as discuss their impacts based on their model architecture and approach categories. Finally, we present challenges and future directions according to technological advancements and the demands of real-world requirements. Although early studies [135], [136], [137] that also survey RRG research from perspectives of different aspects, e.g., its domain and techniques, which mainly categorize RRG studies by comparing RRG with other similar tasks (e.g. image captioning) and put emphasis on particular components (e.g., activation functions) in their models, rather than focusing on the modality differences in methodology settings. Compared to them, this paper offers an updated and systematic review of RRG by focusing on deep learning-based approaches, which is expected to serve as a comprehensive half-decade summary with helpful insights to guide future research in this particular area.

References

[1] F. Narváez, G. Díaz, C. Poveda, and E. Romero, "An Automatic BI-RADS Description of Mammographic Masses by Fusing Multiresolution Features," *Expert Systems with Applications*, vol. 74, pp. 82–95, 2017.

[2] C.-H. Wei, Y. Li, and P. J. Huang, "Mammogram Retrieval through Machine Learning within BI-RADS Standards," *Journal of Biomedical Informatics*, vol. 44, no. 4, pp. 607–614, 2011.

[3] E. Burnside, D. Rubin, and R. Shachter, "A Bayesian Network for Mammography," *AMIA Symposium*, pp. 106–10, 02 2000.

[4] D. L. Rubin, E. S. Burnside, and R. Shachter, *A Bayesian Network to Assist Mammography Interpretation*, Boston, MA, 2004, pp. 695–720.

[5] B. Jing, P. Xie, and E. Xing, "On the Automatic Generation of Medical Imaging Reports," in *ACL 2018*, Melbourne, Australia, Jul. 2018, pp. 2577–2586.

[6] Y. Li, X. Liang, Z. Hu, and E. P. Xing, "Hybrid Retrieval-Generation Reinforced Agent for Medical Image Report Generation," in *NeurIPS 2018*, 2018, pp. 1537–1547.

[7] T. Nishino, R. Ozaki, Y. Momoki, T. Taniguchi, R. Kano, N. Nakano, Y. Tagawa, M. Taniguchi, T. Ohkuma, and K. Nakamura, "Reinforcement Learning with Imbalanced Dataset for Data-to-Text Medical Report Generation," in *Findings of EMNLP 2020*, Online, Nov. 2020, pp. 2223–2236.

[8] Z. Chen, Y. Song, T.-H. Chang, and X. Wan, "Generating Radiology Reports via Memory-driven Transformer," in *EMNLP 2020*, Online, Nov. 2020, pp. 1439–1449.

[9] F. Nooralahzadeh, N. Perez Gonzalez, T. Frauenfelder, K. Fujimoto, and M. Krauthammer, "Progressive Transformer-Based Generation of Radiology Reports," in *Findings of EMNLP 2021*, Punta Cana, Dominican Republic, Nov. 2021, pp. 2824–2832.

[10] J. You, D. Li, M. Okumura, and K. Suzuki, "JPG - Jointly Learn to Align: Automated Disease Prediction and Radiology Report Generation," in *COLING 2022*, Oct. 2022, pp. 5989–6001.

[11] J.-B. Delbrouck, P. Chambon, C. Bluethgen, E. Tsai, O. Almusa, and C. Langlotz, "Improving the Factual Correctness of Radiology Report Generation with Semantic Rewards," in *Findings of EMNLP 2022*, Abu Dhabi, United Arab Emirates, Dec. 2022, pp. 4348–4360.

[12] Z. Wang, L. Liu, L. Wang, and L. Zhou, "METransformer: Radiology Report Generation by Transformer with Multiple Learnable Expert Tokens," in *CVPR 2023*, 2023, pp. 11 558–11 567.

[13] Z. Huang, X. Zhang, and S. Zhang, "KiUT: Knowledge-injected U-Transformer for Radiology Report Generation," in *CVPR 2023*, 2023, pp. 19 809–19 818.

[14] B. Jing, Z. Wang, and E. Xing, "Show, Describe and Conclude: On Exploiting the Structure Information of Chest X-ray Reports," in *ACL 2019*, Florence, Italy, Jul. 2019, pp. 6570–6580.

[15] C. Y. Li, X. Liang, Z. Hu, and E. P. Xing, "Knowledge-driven Encode, Retrieve, Paraphrase for Medical Image Report Generation," in *AAAI 2019*, 2019, pp. 6666–6673.

[16] Y. Zhang, X. Wang, Z. Xu, Q. Yu, A. L. Yuille, and D. Xu, "When Radiology Report Generation Meets Knowledge Graph," in *AAAI 2020*, 2020, pp. 12 910–12 917.

[17] 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*, 2021, pp. 13 753–13 762.

[18] Z. Chen, Y. Shen, Y. Song, and X. Wan, "Cross-modal Memory Networks for Radiology Report Generation," in *ACL-IJCNLP 2021*, Online, Aug. 2021, pp. 5904–5914.

[19] Y. Zhou, L. Huang, T. Zhou, H. Fu, and L. Shao, "Visual-textual Attentive Semantic Consistency for Medical Report Generation," in *ICCV 2021*, 2021, pp. 3965–3974.

[20] T. Nishino, Y. Miura, T. Taniguchi, T. Ohkuma, Y. Suzuki, S. Kido, and N. Tomiyama, "Factual Accuracy is not Enough: Planning Consistent Description Order for Radiology Report Generation," in *EMNLP 2022*, Abu Dhabi, United Arab Emirates, Dec. 2022, pp. 7123–7138.

[21] H. Qin and Y. Song, "Reinforced Cross-modal Alignment for Radiology Report Generation," in *Findings of ACL 2022*, Dublin, Ireland, May 2022, pp. 448–458.

[22] K. Kale, P. Bhattacharyya, and K. Jadhav, "Replace and Report: NLP Assisted Radiology Report Generation," in *Findings of ACL 2023*, Toronto, Canada, Jul. 2023, pp. 10 731–10 742.

[23] T. Tanida, P. Müller, G. Kaissis, and D. Rueckert, "Interactive and Explainable Region-guided Radiology Report Generation," in *CVPR 2023*, 2023, pp. 7433–7442.

[24] M. Li, R. Liu, F. Wang, X. Chang, and X. Liang, "Auxiliary Signal-guided Knowledge Encoder-decoder for Medical Report Generation," *World Wide Web*, pp. 1–18, 2022.

[25] K. Kale, P. Bhattacharyya, M. Gune, A. Shetty, and R. Lawyer, "KGVL-BART: Knowledge Graph Augmented Visual Language BART for Radiology Report Generation," in *EACL 2023*, Dubrovnik, Croatia, May 2023, pp. 3401–3411.

[26] W. Hou, K. Xu, Y. Cheng, W. Li, and J. Liu, "ORGAN: Observation-Guided Radiology Report Generation via Tree Reasoning," in *ACL 2023)*, Toronto, Canada, Jul. 2023, pp. 8108–8122.

[27] A. Yan, Z. He, X. Lu, J. Du, E. Chang, A. Gentili, J. McAuley, and C.-N. Hsu, "Weakly Supervised Contrastive Learning for Chest X-Ray Report Generation," in *Findings of EMNLP 2021*, Punta Cana, Dominican Republic, Nov. 2021, pp. 4009–4015.

[28] Z. Wang, L. Zhou, L. Wang, and X. Li, "A Self-boosting Framework for Automated Radiographic Report Generation," in *CVPR 2021*, 2021, pp. 2433–2442.

[29] F. Liu, C. Yin, X. Wu, S. Ge, P. Zhang, and X. Sun, "Contrastive Attention for Automatic Chest X-ray Report Generation," in *Findings of ACL-IJCNLP 2021*, Online, Aug. 2021, pp. 269–280.

[30] Y. Miura, Y. Zhang, E. Tsai, C. Langlotz, and D. Jurafsky, "Improving Factual Completeness and Consistency of Image-to-text Radiology Report Generation," in *NAACL 2021*, Online, Jun. 2021, pp. 5288–5304.

[31] J. Mao, W. Xu, Y. Yang, J. Wang, and A. L. Yuille, "Deep Captioning with Multimodal Recurrent Neural Networks (m-RNN)," in *ICLR 2015*, 2015, pp. 1–17.

[32] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, "Self-critical Sequence Training for Image Captioning," in *CVPR 2017*. IEEE Computer Society, 2017, pp. 1179–1195.

[33] J. Lu, C. Xiong, D. Parikh, and R. Socher, "Knowing When to Look: Adaptive Attention via a Visual Sentinel for Image Captioning," in *CVPR 2017*, 2017, pp. 3242–3250.

[34] 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.

[35] M. Cornia, M. Stefanini, L. Baraldi, and R. Cucchiara, "Meshed-memory Transformer for Image Captioning," in *CVPR*, 2020, pp. 10 578–10 587.

[36] Y. Wang, K. Wang, X. Liu, T. Gao, J. Zhang, and G. Wang, "Self Adaptive Global-Local Feature Enhancement for Radiology Report Generation," *2023 IEEE International Conference on Image Processing (ICIP)*, pp. 2275–2279, 2022.

[37] 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*, 2018, pp. 457–466.

[38] J. Yuan, H. Liao, R. Luo, and J. Luo, "Automatic Radiology Report Generation based on Multi-view Image Fusion and Medical Concept Enrichment," *ArXiv*, vol. abs/1907.09085, 2019.

[39] P. Harzig, Y. Chen, F. Chen, and R. Lienhart, "Addressing Data Bias Problems for Chest X-ray Image Report Generation," in *BMVC 2019*, 2019, p. 144.

[40] F. Liu, C. You, X. Wu, S. Ge, X. Sun *et al.*, "Auto-encoding Knowledge Graph for Unsupervised Medical Report Generation," *Advances in Neural Information Processing Systems*, vol. 34, pp. 16 266–16 279, 2021.

[41] F. Dalla Serra, W. Clackett, H. MacKinnon, C. Wang, F. Deligianni, J. Dalton, and A. Q. O'Neil, "Multimodal Generation of Radiology Reports using Knowledge-Grounded Extraction of Entities and Relations," in *AACL 2022*, Online only, Nov. 2022, pp. 615–624.

[42] S. Yan, W. K. Cheung, W. H. K. Chiu, T. M. Tong, K. C. Cheung, and S. See, "Attributed Abnormality Graph Embedding for Clinically Accurate X-Ray Report Generation," *IEEE Trans. Medical Imaging*, vol. 42, no. 8, pp. 2211–2222, 2023.

[43] M. Li, B. Lin, Z. Chen, H. Lin, X. Liang, and X. Chang, "Dynamic Graph Enhanced Contrastive Learning for Chest X-Ray Report Generation," in *CVPR 2023*, Los Alamitos, CA, USA, jun 2023, pp. 3334–3343.

[44] K. Zhang, H. Jiang, J. Zhang, Q. Huang, J. Fan, J. Yu, and W. Han, "Semi-supervised Medical Report Generation via Graph-guided Hybrid Feature Consistency," *IEEE Transactions on Multimedia*, pp. 1–13, 2023.

[45] B. Hou, G. Kaissis, R. M. Summers, and B. Kainz, "RATCHET: Medical Transformer for Chest X-ray Diagnosis and Reporting," in *MICCAI 2021*, vol. 12907, 2021, pp. 293–303.

[46] 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*, vol. 12903, 2021, pp. 72–82.

[47] J. Wang, A. Bhalerao, and Y. He, "Cross-modal prototype driven network for radiology report generation," in *ECCV*, vol. 13695, 2022, pp. 563–579.

[48] H. Yu and Q. Zhang, "Clinically Coherent Radiology Report Generation with Imbalanced Chest X-rays," in *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2022, pp. 1781–1786.

[49] B. Yan, M. Pei, M. Zhao, C. Shan, and Z. Tian, "Prior Guided Transformer for Accurate Radiology Reports Generation," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 11, pp. 5631–5640, 2022.

[50] L. Wang, M. Ning, D. Lu, D. Wei, Y. Zheng, and J. Chen, "An Inclusive Task-Aware Framework for Radiology Report Generation," in *MICCAI 2022*, vol. 13438, 2022, pp. 568–577.

[51] Z. Wang, M. Tang, L. Wang, X. Li, and L. Zhou, "A Medical Semantic-Assisted Transformer for Radiographic Report Generation," in *MICCAI 2022*, vol. 13433, 2022, pp. 655–664.

[52] M. Kong, Z. Huang, K. Kuang, Q. Zhu, and F. Wu, "TranSQ: Transformer-Based Semantic Query for Medical Report Generation," in *MICCAI 2022*, vol. 13438, 2022, pp. 610–620.

[53] A. K. Tanwani, J. K. Barral, and D. Freedman, "Repsnet: Combining vision with language for automated medical reports," in *MICCAI 2022*, vol. 13435, 2022, pp. 714–724.

[54] Z. Wang, H. Han, L. Wang, X. Li, and L. Zhou, "Automated Radiographic Report Generation Purely on Transformer: A Multicriteria Supervised Approach," *IEEE Transactions on Medical Imaging*, vol. 41, no. 10, pp. 2803–2813, 2022.

[55] Y. Yang, J. Yu, J. Zhang, W. Han, H. Jiang, and Q. Huang, "Joint Embedding of Deep Visual and Semantic Features for Medical Image Report Generation," *IEEE Transactions on Multimedia*, vol. 25, pp. 167–178, 2023.

[56] A. Nicolson, J. Dowling, and B. Koopman, "Improving Chest X-ray Report Generation by Leveraging Warm Starting," *Artificial Intelligence in Medicine*, vol. 144, p. 102633, 2023.

[57] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-scale Hierarchical Image Database," in *CVPR 2009*. Ieee, 2009, pp. 248–255.

[58] T. Tanida, P. Müller, G. Kaissis, and D. Rueckert, "Interactive and Explainable Region-guided Radiology Report Generation," in *CVPR*, 2023.

[59] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," in *ICLR 2015*, 2015, pp. 1–14.

[60] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition*, ser. CVPR '16, Jun. 2016, pp. 770–778.

[61] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," in *CVPR*, 2017, pp. 2261–2269.

[62] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is All You Need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[63] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in *ICLR*, 2021, pp. 1–21.

[64] J. Uijlings, K. Sande, T. Gevers, and A. Smeulders, "Selective Search for Object Recognition," *International Journal of Computer Vision*, vol. 104, pp. 154–171, 09 2013.

[65] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," *TPAMI*, vol. 39, no. 6, pp. 1137–1149, 2017.

[66] J. T. Wu, N. Agu, I. Lourentzou, A. Sharma, J. A. Paguio, J. S. Yao, E. C. Dee, W. Mitchell, S. Kashyap, A. Giovannini, L. A. Celi, and M. Moradi, "Chest ImaGenome Dataset for Clinical Reasoning," in *NeurIPS Datasets and Benchmarks 2021*, 2021, pp. 1–14.

[67] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[68] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. L. Ball, K. S. Shpanskaya, J. Seekins, D. A. Mong, S. S. Halabi, J. K. Sandberg, R. Jones, D. B. Larson, C. P. Langlotz, B. N. Patel, M. P. Lungren, and A. Y. Ng, "CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison," in *AAAI 2019*, 2019, pp. 590–597.

[69] S. Jain, A. Smit, S. Q. Truong, C. D. Nguyen, M.-T. Huynh, M. Jain, V. A. Young, A. Y. Ng, M. P. Lungren, and P. Rajpurkar, "VisualCheXBERT: Addressing the Discrepancy between Radiology Report Labels and Image Labels," ser. CHIL '21, New York, NY, USA, 2021, p. 105–115.

[70] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the Limits of Transfer Learning with a Unified Text-to-text Transformer," *Journal of Machine Learning Research*, vol. 21, no. 140, pp. 1–67, 2020.

[71] S. Jain, A. Agrawal, A. Saporta, S. Truong, D. N. D. N. Duong, T. Bui, P. Chambon, Y. Zhang, M. Lungren, A. Ng, C. Langlotz, P. Rajpurkar, and P. Rajpurkar, "RadGraph: Extracting Clinical Entities and Relations from Radiology Reports," in *NeurIPS*, vol. 1, 2021, pp. 1–12.

[72] M. Neumann, D. King, I. Beltagy, and W. Ammar, "ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing," in *BioNLP*, Florence, Italy, Aug. 2019, pp. 319–327.

[73] T. N. Kipf and M. Welling, "Semi-Supervised Classification with Graph Convolutional Networks," in *International Conference on Learning Representations*, 2017, pp. 1–14.

[74] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *ACL 2020*, Online, Jul. 2020, pp. 7871–7880.

[75] K. Kale, P. Bhattacharyya, A. Shetty, M. Gune, K. Shrivastava, R. Lawyer, and S. Biswas, "Knowledge Graph Construction and Its Application in Automatic Radiology Report Generation from Radiologist's Dictation," *CoRR*, vol. abs/2206.06308, 2022.

[76] P. Qi, Y. Zhang, Y. Zhang, J. Bolton, and C. D. Manning, "Stanza: A python natural language processing toolkit for many human languages," in *ACL: System Demonstrations*, Online, Jul. 2020, pp. 101–108.

[77] A. Smit, S. Jain, P. Rajpurkar, A. Pareek, A. Ng, and M. Lungren, "Combining Automatic Labelers and Expert Annotations for Accurate Radiology Report Labeling Using BERT," in *EMNLP 2020*, Online, Nov. 2020, pp. 1500–1519.

[78] F. Liu, S. Ge, and X. Wu, "Competence-based Multimodal Curriculum Learning for Medical Report Generation," in *ACL-IJCNLP 2021*, Online, Aug. 2021, pp. 3001–3012.

[79] M. Ranzato, S. Chopra, M. Auli, and W. Zaremba, "Sequence Level Training with Recurrent Neural Networks," in *ICLR*, 2016.

[80] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A Method for Automatic Evaluation of Machine Translation," in *ACL*, Philadelphia, Pennsylvania, USA, Jul. 2002, pp. 311–318.

[81] S. Banerjee and A. Lavie, "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments," in *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, Ann Arbor, Michigan, Jun. 2005, pp. 65–72.

[82] C.-Y. Lin, "ROUGE: A Package for Automatic Evaluation of Summaries," in *Text Summarization Branches Out*, Barcelona, Spain, Jul. 2004, pp. 74–81.

[83] R. J. Williams, "Simple Statistical Gradient-Following Algorithms for Connectionist Reinforcement Learning," *Mach. Learn.*, vol. 8, pp. 229–256, 1992.

[84] R. Vedantam, C. L. Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *CVPR 2015*, 2015, pp. 4566–4575.

[85] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *NAACL 2019)*, Minneapolis, Minnesota, Jun. 2019, pp. 4171–4186.

[86] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum Learning," in *ICML*, New York, NY, USA, 2009, p. 41–48.

[87] F. Liu, X. Ren, Y. Liu, K. Lei, and X. Sun, "Exploring and Distilling Cross-Modal Information for Image Captioning," in *IJCAI 2019*, 2019, pp. 5095–5101.

[88] X. Zhang, G. Kumar, H. Khayrallah, K. Murray, J. Gwinnup, M. J. Martindale, P. McNamee, K. Duh, and M. Carpuat, "An Empirical Exploration of Curriculum Learning for Neural Machine Translation," *CoRR*, vol. abs/1811.00739, 2018.

[89] F. Faghri, D. J. Fleet, J. R. Kiros, and S. Fidler, "VSE++: Improving Visual-Semantic Embeddings with Hard Negatives," in *BMVC 2018*, 2018, p. 12.

[90] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative Adversarial Nets," in *Advances in Neural Information Processing Systems*, vol. 27, 2014, pp. 1–9.

[91] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, and H. Poon, "Domain-specific Language Model Pretraining for Biomedical Natural Language Processing," *ACM Trans. Comput. Heal.*, vol. 3, no. 1, pp. 2:1–2:23, 2022.

[92] C. Raffel and D. P. W. Ellis, "Feed-forward networks with attention can solve some long-term memory problems," *CoRR*, vol. abs/1512.08756, 2015.

[93] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and Tell: A Neural Image Caption Generator," in *CVPR 2015*, 2015.

[94] D. Demner-Fushman, M. D. Kohli, M. B. Rosenman, S. E. Shooshan, L. Rodriguez, S. K. Antani, G. R. Thoma, and C. J. McDonald, "Preparing A Collection of Radiology Examinations for Distribution and Retrieval," *J. Am. Medical Informatics Assoc.*, vol. 23, no. 2, pp. 304–310, 2016.

[95] A. E. Johnson, T. J. Pollard, S. J. Berkowitz, N. R. Greenbaum, M. P. Lungren, C.-y. Deng, R. G. Mark, and S. Horng, "MIMIC-CXR: A De-identified Publicly Available Database of Chest Radiographs with Free-text Reports," *Scientific Data*, vol. 6, 2019.

[96] A. E. W. Johnson, T. J. Pollard, S. J. Berkowitz, N. R. Greenbaum, M. P. Lungren, C. Deng, R. G. Mark, and S. Horng, "MIMIC-CXR: A Large Publicly Available Database of Labeled Chest Radiographs," *CoRR*, vol. abs/1901.07042, 2019.

[97] J. Ni, C. Hsu, A. Gentili, and J. J. McAuley, "Learning Visual-Semantic Embeddings for Reporting Abnormal Findings on Chest X-rays," in *Findings of EMNLP*, vol. EMNLP 2020, 2020, pp. 1954–1960.

[98] S. Sukhbaatar, a. szlam, J. Weston, and R. Fergus, "End-to-end Memory Networks," in *Advances in Neural Information Processing Systems*, vol. 28, 2015, pp. 1–9.

[99] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "ChestX-Ray8: Hospital-scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases," in *CVPR 2017*, 2017, pp. 3462–3471.

[100] J. Liu, J. Lian, and Y. Yu, "ChestX-Det10: Chest X-ray Dataset on Detection of Thoracic Abnormalities," *CoRR*, vol. abs/2006.10550, 2020.

[101] 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*, 2018, pp. 9049–9058.

[102] Z.-M. Chen, X.-S. Wei, P. Wang, and Y. Guo, "Multi-label image recognition with graph convolutional networks," in *CVPR 2019*, 2019, pp. 5172–5181.

[103] B. Chen, Y. Lu, and G. Lu, "Multi-label Chest X-ray Image Classification via Label Co-occurrence Learning," in *PRCV 2019*, Berlin, Heidelberg, 2019, p. 682–693.

[104] A. Smit, S. Jain, P. Rajpurkar, A. Pareek, A. Y. Ng, and M. P. Lungren, "CheXBERT: Combining Automatic Labelers and Expert Annotations for Accurate Radiology Report Labeling Using BERT," *CoRR*, vol. abs/2004.09167, 2020.

[105] Y. Zhang and R. Jiao, "How Segment Anything Model (SAM) Boost Medical Image Segmentation?" *arXiv preprint arXiv:2305.03678*, 2023.

[106] K. Zhang and D. Liu, "Customized Segment Anything Model for Medical Image Segmentation," *arXiv preprint arXiv:2304.13785*, 2023.

[107] D. Anand, G. R. M, V. Singhal, D. D. Shanbhag, K. S. Shriram, U. Patil, C. Bhushan, K. Manickam, D. Gui, R. Mullick, A. Gopal, P. Bhatia, and T. A. Kass-Hout, "One-shot Localization and Segmentation of Medical Images with Foundation Models," *CoRR*, vol. abs/2310.18642, 2023.

[108] A. Ranem, N. Babendererde, M. Fuchs, and A. Mukhopadhyay, "Exploring SAM Ablations for Enhancing Medical Segmentation in Radiology and Pathology," 2023.

[109] S. Pandey, K.-F. Chen, and E. B. Dam, "Comprehensive Multimodal Segmentation in Medical Imaging: Combining YOLOv8 with SAM and HQ-SAM Models," in *ICCV Workshops*, October 2023, pp. 2592–2598.

[110] C. Wang, X. Chen, H. Ning, and S. Li, "SAM-OCTA: A Fine-Tuning Strategy for Applying Foundation Model to OCTA Image Segmentation Tasks," 2023.

[111] P. Zhang and Y. Wang, "Segment Anything Model for Brain Tumor Segmentation," 2023.

[112] B. Fazekas, J. Morano, D. Lachinov, G. Aresta, and H. Bogunović, "Adapting Segment Anything Model (SAM) For Retinal OCT," in *MICCAI 2023*, Berlin, Heidelberg, 2023, p. 92–101.

[113] N. Wang, Y. Song, and F. Xia, "Coding structures and actions with the COSTA scheme in medical conversations," in *BioNLP*, Melbourne, Australia, Jul. 2018, pp. 76–86.

[114] Y. Tian, W. Ma, F. Xia, and Y. Song, "ChiMed: A Chinese Medical Corpus for Question Answering," in *BioNLP*, Florence, Italy, Aug. 2019, pp. 250–260.

[115] N. Wang, Y. Song, and F. Xia, "Studying challenges in medical conversation with structured annotation," in *NLPMC*, Online, Jul. 2020, pp. 12–21.

[116] Y. Song, Y. Tian, N. Wang, and F. Xia, "Summarizing Medical Conversations via Identifying Important Utterances," in *Proceedings of the 28th International Conference on Computational Linguistics*, Barcelona, Spain (Online), Dec. 2020, pp. 717–729.

[117] K. Krishna, S. Khosla, J. Bigham, and Z. C. Lipton, "Generating SOAP notes from doctor-patient conversations using modular summarization techniques," in *ACL 2021*, Online, Aug. 2021, pp. 4958–4972.

[118] G. Michalopoulos, K. Williams, G. Singh, and T. Lin, "MedicalSum: A guided clinical abstractive summarization model for generating medical reports from patient-doctor conversations," in *Findings of EMNLP 2022*, Abu Dhabi, United Arab Emirates, Dec. 2022, pp. 4741–4749.

[119] Y. Peng, X. Wang, L. Lu, M. Bagheri, R. M. Summers, and Z. Lu, "NegBio: A High-performance Tool for Negation and Uncertainty Detection in Radiology Reports," *CoRR*, vol. abs/1712.05898, 2017.

[120] T. Zhang*, V. Kishore*, F. Wu*, K. Q. Weinberger, and Y. Artzi, "BERTScore: Evaluating Text Generation with BERT," in *ICLR*, 2020, pp. 1–43.

[121] J. Zhao, Y. Zhang, X. He, and P. Xie, "COVID-CT-Dataset: A CT Scan Dataset about COVID-19," *arXiv preprint arXiv:2003.13865*, 2020.

[122] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollar, and R. Girshick, "Segment Anything," in *ICCV*, October 2023, pp. 4015–4026.

[123] J. Ho, A. Jain, and P. Abbeel, "Denoising Diffusion Probabilistic Models," *NeurIPS*, vol. 33, pp. 6840–6851, 2020.

[124] S. Gong, M. Li, J. Feng, Z. Wu, and L. Kong, "DiffuSeq: Sequence to Sequence Text Generation with Diffusion Models," in *ICLR*, 2023, pp. 1–20.

[125] J. Luo, Y. Li, Y. Pan, T. Yao, J. Feng, H. Chao, and T. Mei, "Semantic-conditional Diffusion Networks for Image Captioning," in *CVPR 2023*, 2023, pp. 23 359–23 368.

[126] T. Chen, R. Zhang, and G. Hinton, "Analog Bits: Generating Discrete Data using Diffusion Models with Self-conditioning," in *ICLR*, 2023, pp. 1–23.

[127] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample, "LLaMA: Open and Efficient Foundation Language Models," *CoRR*, vol. abs/2302.13971, 2023.

[128] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. Canton-Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, and T. Scialom, "Llama 2: Open Foundation and Fine-Tuned Chat Models," *CoRR*, vol. abs/2307.09288, 2023.

[129] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny, "MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models," *CoRR*, vol. abs/2304.10592, 2023.

[130] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual Instruction Tuning," *CoRR*, vol. abs/2304.08485, 2023.

[131] Q. Ye, H. Xu, G. Xu, J. Ye, M. Yan, Y. Zhou, J. Wang, A. Hu, P. Shi, Y. Shi, C. Li, Y. Xu, H. Chen, J. Tian, Q. Qi, J. Zhang, and F. Huang, "mPLUG-Owl: Modularization Empowers Large Language Models with Multimodality," *CoRR*, vol. abs/2304.14178, 2023.

[132] H. Xu, Q. Ye, M. Yan, Y. Shi, J. Ye, Y. Xu, C. Li, B. Bi, Q. Qian, W. Wang, G. Xu, J. Zhang, S. Huang, F. Huang, and J. Zhou, "mPLUG-2: A Modularized Multi-modal Foundation Model Across Text, Image and Video," 2023.

[133] O. Thawakar, A. Shaker, S. S. Mullappilly, H. Cholakkal, R. M. Anwer, S. H. Khan, J. Laaksonen, and F. S. Khan, "XrayGPT: Chest Radiographs Summarization using Medical Vision-Language Models," *CoRR*, vol. abs/2306.07971, 2023.

[134] J. Hessel, A. Holtzman, M. Forbes, R. Le Bras, and Y. Choi, "CLIPScore: A Reference-free Evaluation Metric for Image Captioning," in *EMNLP 2021*, Online and Punta Cana, Dominican Republic, Nov. 2021, pp. 7514–7528.

[135] J. Pavlopoulos, V. Kougia, and I. Androutsopoulos, "A Survey on Biomedical Image Captioning," in *Proceedings of the Second Workshop on Shortcomings in Vision and Language*, Minneapolis, Minnesota, Jun. 2019, pp. 26–36.

[136] M. M. A. Monshi, J. Poon, and V. Chung, "Deep Learning in Generating Radiology Reports: A Survey," *Artificial Intelligence in Medicine*, vol. 106, p. 101878, 2020.

[137] J. Pavlopoulos, V. Kougia, I. Androutsopoulos, and D. Papamichail, "Diagnostic Captioning: A Survey," *Knowledge and Information Systems*, vol. 64, pp. 1–32, 07 2022.